# A SCALABLE ALGORITHM FOR GAUSSIAN GRAPHICAL MODELS WITH CHANGE-POINTS

LELAND BYBEE AND YVES ATCHADÉ

(March 2018; First version March 2017)

Abstract. Graphical models with change-points are computationally challenging to fit, particularly in cases where the number of observation points and the number of nodes in the graph are large. Focusing on Gaussian graphical models, we introduce an approximate majorize-minimize (MM) algorithm that can be useful for computing change-points in large graphical models. The proposed algorithm is an order of magnitude faster than a brute force search. Under some regularity conditions on the data generating process, we show that with high probability, the algorithm converges to a value that is within statistical error of the true change-point. A fast implementation of the algorithm using Markov Chain Monte Carlo is also introduced. The performances of the proposed algorithms are evaluated on synthetic data sets and the algorithm is also used to analyze structural changes in the S&P 500 over the period 2000-2016.

## 1. Introduction

Networks are fundamental structures that are commonly used to describe interactions between sets of actors or nodes. In many applications, the behaviors of the actors are observed over time and one is interested in recovering the underlying network connecting these actors. High-dimensional versions of this problem where the number of actors is large (compared to the number of time points) is of special interest. In the statistics and machine learning literature, this problem is typically framed as fitting large graphical models with sparse parameters, and significant progress has been made recently, both in terms of the statistical theory (Meinshausen and Buhlmann (2006); Yuan and Lin (2007); Banerjee et al. (2008); Ravikumar et al. (2011); Hastie

2010 *Mathematics Subject Classification.* 62F15, 62Jxx.

*Key words and phrases.* Change-points, Gaussian graphical models, proximal gradient, simulated Annealing, Stochastic Optimization.

This work is partially supported by the NSF grant DMS 1513040.

L. Bybee: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* lelandb@umich.edu.

Y. Atchadé: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* yvesa@umich.edu.

et al. (2015)), and practical algorithms (Friedman et al. (2007); Höfling and Tibshirani (2009); Atchade et al. (2017)).

In many problems arising in areas such as biology, finance, and political sciences, it is well-accepted that the underlying networks of interest are not static, but can undergo changes over time. Graphical models with change-points (or piecewise constant graphical models) are simple, yet powerful models that are particularly well-suited for such problems, and different versions have been explored in the literature. In this work, similarly to Zhou et al. (2009); Kolar et al. (2010); Roy et al. (2017), we focus on settings where the change occurring at a given change-point is global in the sense that it affects the joint distribution of all nodes. This differs from the approach of Kolar and Xing (2012) where at a given change-point only the conditional distribution of a single node sees a change. Which framework is more appropriate depends in general on the application. For instance in biological applications where interests are often on single biomolecules, nodewise change-point analysis might be preferred, whereas in many social science problems global structural changes in the network is often of interest. We also mention the alternative approach of Liu et al. (2013) which has an original parametrization that focuses directly on the occurring change. Although we work within the joint-change framework, we stress that our proposed algorithms can be easily adapted to work with other alternative models.

Despite their conceptual simplicity, graphical models with change-points are computationally challenging to fit. For instance a full grid search approach to locate a single change-point in a Gaussian graphical model with a lasso penalty (glasso) requires solving $O(T)$ glasso sub-problems, where $T$ is the number of time points. Most algorithms for the glasso problem scale like $O(p^3)$ or worst[1], where $p$ is the number of nodes. Hence when $p$ and $T$ are large, fitting a high-dimensional Gaussian graphical model with a single change-point has a taxing computational cost of $O(Tp^3)$ per iteration.

The literature addressing the computational aspects of model-based change-point models is rather sparse. A large portion of change-point detection procedures are based on cumulative sums (CUSUM) or similar statistic-monitoring approaches (Lévy-Leduc and Roueff (2009); Aue et al. (2009); Fryzlewicz (2014); Chen and Zhang (2015); Cho and Fryzlewicz (2015) and the references therein). By and large, these change-point detection procedures can be efficiently implemented, and the computational difficulty aforementioned can be avoided. However in problems where one wishes to detect structural changes in large networks, a CUSUM-based or a statistic-based approach can be difficult to employ, since it requires knowledge of the pertinent

[1]Furthermore the constant in the big-O is typically problem dependent and can be large

statistics to monitor. Furthermore the estimation of the parameters in a model-based change-point models can provide new insight in the underlying phenomenon driving the changes. Hence CUSUM-based approaches may not be appropriate in applications where the main driving forces of the network changes are poorly understood, and/or are of prime interest.

Specific works addressing computational issues in model-based change-point estimation include Roy et al. (2017); Leonardi and Bühlmann (2016). In Roy et al. (2017) the authors considered a discrete graphical model with change-point and proposed a two-steps algorithm for computation. However the success of their algorithm depends crucially on the choice of the coarse and refined grids, and there is limited insight on how to choose these. A related work is Leonardi and Bühlmann (2016) where the authors considered a high-dimensional linear regression model with change-points and proposed a dynamic programming approach to compute the change points. In the case of a single change-point their algorithm corresponds to the brute force (full-grid search) approach mentioned above.

In this work we propose an approximate majorize-minimize (MM) algorithm for fitting piecewise constant high-dimensional models. The algorithm can be applied more broadly. However to focus the idea we limit our discuss to Gaussian graphical models with an elastic net penalty. In this specific setting, the algorithm takes the form of a block update algorithm that alternates between a proximal gradient update of the graphical model parameters followed by a line search of the change-point. The proposed algorithm only solves for a single change-point. We extend it to multiple change-points by binary segmentation. We study the convergence of the algorithm and show under some regularity conditions on the data generating mechanism that the algorithm is stable, and produces values in the vicinity of the true change-point (under the assumption that one such true change-point exists).

Each iteration of the proposed algorithm has a computational cost of $O(Tp^2 + p^3)$. Although this cost is one order of magnitude smaller than the $O(Tp^3)$ cost of the brute force approach, it can still be large when $p$ and $T$ are both large. As a solution we propose a stochastic version of the algorithm where the line search performed to update the change-point is replaced by a Markov Chain Monte Carlo (MCMC)-based simulated annealing. The simulated annealing update is cheap (its computational cost per iteration is $O(p^2)$) and is used as a stochastic approximation of the full line search. We show by simulation that the stochastic algorithm behaves remarkably well, and as expected outperforms the deterministic algorithm is terms of computing time.

The paper is organized as follows. Section 2 contains a presentation of the Gaussian graphical model with change-points, followed by a detailed presentation of the proposed algorithms. We performed extensive numerical experiments to investigate the behavior of the proposed algorithms. We also use the algorithm to analyze structural changes in the Standard & Poors (S&P) 500 over the period 2000-2016. The results are reported in Section 3. We gather some of the technical proofs in Section 4.

We end this introduction with some notation that we shall use throughout the paper. We denote $\mathcal{M}_p$ the set of all symmetric elements of $\mathbb{R}^{p\times p}$ equipped with its Frobenius norm $\|\cdot\|_{\mathsf{F}}$ and associated inner product

$$\langle A, B\rangle_{\mathsf{F}} \stackrel{\text{def}}{=} \sum_{1\le i\le j\le p} A_{ij}B_{ij}.$$

We denote $\mathcal{M}_p^+$ the subset of $\mathcal{M}_p$ of positive definite elements. For $0 < a < A \le +\infty$, let $\mathcal{M}_p^+(a, A)$ denote the subset of $\mathcal{M}_p^+$ of matrices $\theta$ such that $\lambda_{\mathsf{min}}(\theta) \ge a$, and $\lambda_{\mathsf{max}}(\theta) \le A$, where $\lambda_{\mathsf{min}}(M)$ (resp. $\lambda_{\mathsf{max}}(M)$) denotes the smallest eigenvalue (resp. the largest eigenvalue) of $M$.

If $u \in \mathbb{R}^p$, and $q \in [1, \infty]$, we define $\|u\|_q \stackrel{\text{def}}{=} (\sum_{j=1}^p |u_j|^q)^{1/q}$ ($\|u\|_\infty \stackrel{\text{def}}{=} \max_{1\le j\le p} |u_j|$). For a matrix $\theta \in \mathbb{R}^{p\times p}$ and $q \in [1, \infty] \setminus \{2\}$, we define $\|\theta\|_q$ similarly by viewing $\theta$ as a $\mathbb{R}^{p^2}$ vector. For $q = 2$, $\|\theta\|_2$ denotes the spectral norm (operator norm) of $\theta$.

## 2. Fitting Gaussian Graphical models with a single change-point

Let $\{X^{(t)},\ 1 \le t \le T\}$ be a sequence of $p$-dimensional random vectors. The grid over which the change-points are searched is denoted $\mathcal{T} \stackrel{\text{def}}{=} \{n_0, \ldots, T - n_0\}$, for some integer $1 \le n_0 < T$. We define

$$S_1(\tau) \stackrel{\text{def}}{=} \frac{1}{\tau}\sum_{t=1}^{\tau} X^{(t)}X^{(t)'}, \quad S_2(\tau) \stackrel{\text{def}}{=} \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} X^{(t)}X^{(t)'}, \ \ \tau \in \mathcal{T}.$$

We define the regularization function as

$$\wp(\theta) \stackrel{\text{def}}{=} \alpha\|\theta\|_1 + \frac{1-\alpha}{2}\|\theta\|_{\mathsf{F}}^2, \quad \theta \in \mathcal{M}_p, \tag{1}$$

where $\alpha \in [0, 1)$ is a given constant, and $\|\theta\|_1 \stackrel{\text{def}}{=} \sum_{i\le j}^p |\theta_{ij}|$. Then we define

$$g_{1,\tau}(\theta) = \begin{cases} \frac{1}{2}\frac{\tau}{T}\left[-\log\det(\theta) + \mathsf{Tr}(\theta S_1(\tau))\right] & \text{if } \theta \in \mathcal{M}_p^+, \\ +\infty & \text{otherwise,} \end{cases}, \ \ \tau \in \mathcal{T},$$

where $\mathsf{Tr}(A)$ (resp. $\det(A)$) denotes the trace (resp. the determinant) of $A$, and

$$g_{2,\tau}(\theta) = \begin{cases} \frac{1}{2}\left(1-\frac{\tau}{T}\right)\left[-\log\det(\theta) + \mathsf{Tr}(\theta S_2(\tau))\right] & \text{if } \theta \in \mathcal{M}_p^+, \\ +\infty & \text{otherwise,} \end{cases}, \ \ \tau \in \mathcal{T}.$$

For $j \in \{1, 2\}$, we set

$$\hat{\theta}_{j,\tau} \stackrel{\text{def}}{=} \mathsf{Argmin}_{\vartheta \in \mathcal{M}_p^+} \left[ g_{j,\tau}(\vartheta) + \lambda_{j,\tau} \wp(\vartheta) \right], \tag{2}$$

for regularization parameters $\lambda_{1,\tau} > 0, \lambda_{2,\tau} > 0$, that we assume fixed throughout. Note that due to the quadratic term in the elastic-net regularization (1), each of these minimization problems (2) is strongly convex. Hence for each $\tau \in \mathcal{T}$, and $j \in \{1, 2\}$, $\hat{\theta}_{j,\tau}$ is well-defined. We consider the problem of computing the change point estimate $\hat{\tau}$ defined as

$$\hat{\tau} = \mathsf{Argmin}_{\tau \in \mathcal{T}} \left[ g_{1,\tau}(\hat{\theta}_{1,\tau}) + \lambda_{1,\tau} \wp(\hat{\theta}_{1,\tau}) + g_{2,\tau}(\hat{\theta}_{2,\tau}) + \lambda_{2,\tau} \wp(\hat{\theta}_{2,\tau}) \right]. \tag{3}$$

If the minimization problem in (3) has more than one solution, then $\hat{\tau}$ denotes any one of these solutions. The quantity $\hat{\tau}$ is the maximum likelihood estimate of a change point $\tau$ in the model which assumes that $X^{(1)}, \ldots, X^{(\tau)}$ are independent with common distribution $\mathbf{N}(0, \theta_1^{-1})$, and $X^{(\tau+1)}, \ldots, X^{(T)}$ are independent with common distribution $\mathbf{N}(0, \theta_2^{-1})$, for an unknown change-point $\tau$, and unknown precision matrices $\theta_1 \neq \theta_2$.

The problem of computing the graphical lasso ($\mathsf{glasso}$) estimators $\hat{\theta}_{j,\tau}$ in (2) has received a lot of attention in the literature, and several efficient algorithms have been developed for this purpose (see for instance Atchadé et al. (2015) and the references therein). Hence in principle, using any of these available $\mathsf{glasso}$ algorithms, the change-point problem in (3) can be solved by solving $T - 2n_0 + 1 = O(T)$ $\mathsf{glasso}$ sub-problems. A similar algorithm is advocated in Leonardi and Bühlmann (2016) for fitting a high-dimensional linear regression model with change-points. However this brute force approach can be very time-consuming in cases where $p$ and $T$ are large. For instance, one of the most cost-efficient algorithm for solving the $\mathsf{glasso}$ problem in high-dimensional cases is the standard proximal gradient algorithm (Rolfs et al. (2012); Atchadé et al. (2015)), which has a computational cost of $O(p^3 \mathsf{cond}(\hat{\theta})^2 \log(1/\delta))$ to deliver a $\delta$-accurate solution (that is $\|\theta - \hat{\theta}\|_{\mathsf{F}} \leq \delta$), where $\mathsf{cond}(A)$ denotes the condition number of $A$, that is the ratio of the largest eigenvalue over the smallest eigenvalue of $A$. Hence when $p$ and $T$ are large the computational cost of the brute force approach for computing (3) is of order $O\left(Tp^3 \mathsf{cond}(\hat{\theta}_{j,\tau})^2 \log(1/\delta)\right)$, which can become prohibitively large.

We propose an algorithm that we show has a better computational complexity. To motivate the algorithm we first introduce a majorize-minimize (MM) algorithm for solving (3). We refer the reader to Wu and Lange (2010) for a general introduction to MM algorithms. Let

$$G(t) \stackrel{\text{def}}{=} g_{1,t}(\hat{\theta}_{1,t}) + \lambda_{1,t} \wp(\hat{\theta}_{1,t}) + g_{2,t}(\hat{\theta}_{2,t}) + \lambda_{2,\tau} \wp(\hat{\theta}_{2,t}), \quad t \in \mathcal{T}$$

denote the objective function of the minimization problem in (3). For $\theta_1, \theta_2 \in \mathcal{M}_p$, we also define

$$\mathcal{H}(\tau|\theta_1,\theta_2) \stackrel{\text{def}}{=} g_{1,\tau}(\theta_1) + \lambda_{1,\tau}\wp(\theta_1) + g_{2,\tau}(\theta_2) + \lambda_{2,\tau}\wp(\theta_2), \;\; \tau \in \mathcal{T}. \tag{4}$$

Instead of the brute force approach that requires solving (2) for each value $\tau \in \mathcal{T}$, consider the following algorithm.

**Algorithm 1** (MM algorithm)**.** Pick $\tau^{(0)} \in \mathcal{T}$, and for $k = 1, \ldots, K$, repeat the following steps.

(1) Given $\tau^{(k-1)} \in \mathcal{T}$, compute $\hat{\theta}_{1,\tau^{(k-1)}}$ and $\hat{\theta}_{2,\tau^{(k-1)}}$, and minimize the function $\mathcal{H}(t|\hat{\theta}_{1,\tau^{(k-1)}}, \hat{\theta}_{2,\tau^{(k-1)}})$ to get $\tau^{(k)}$:

$$\tau^{(k)} = \mathsf{Argmin}_{t\in\mathcal{T}}\, \mathcal{H}(t|\hat{\theta}_{1,\tau^{(k-1)}}, \hat{\theta}_{2,\tau^{(k-1)}}).$$

□

By definition of $\hat{\theta}_{j,\tau}$ in (2), we have $G(t) \leq \mathcal{H}(t|\hat{\theta}_{1,\tau^{(k-1)}}, \hat{\theta}_{2,\tau^{(k-1)}})$ for all $t \in \mathcal{T}$. Furthermore $G(\tau^{(k-1)}) = \mathcal{H}(\tau^{(k-1)}|\hat{\theta}_{1,\tau^{(k-1)}}, \hat{\theta}_{2,\tau^{(k-1)}})$. Therefore, for all $k \geq 1$,

$$G(\tau^{(k)}) \leq \mathcal{H}(\tau^{(k)}|\hat{\theta}_{1,\tau^{(k-1)}}, \hat{\theta}_{2,\tau^{(k-1)}}) \leq \mathcal{H}(\tau^{(k-1)}|\hat{\theta}_{1,\tau^{(k-1)}}, \hat{\theta}_{2,\tau^{(k-1)}}) = G(\tau^{(k-1)}).$$

Hence the objective function $G$ is non-increasing along the iterates of Algorithm 1. Note that this algorithm is already potentially faster than the brute force approach, particular when $T$ is large, since we compute the graphical-lasso solutions $\hat{\theta}_{j,\tau^{(k)}}$ only for time points visited along the iterations. We propose to further reduce the computational cost by computing the solutions $\hat{\theta}_{j,\tau^{(k)}}$ only approximately, by simple gradient updates.

Given $\gamma > 0$, and a matrix $\theta \in \mathbb{R}^{p\times p}$, define $\mathrm{Prox}_\gamma(\theta)$ (the proximal map with respect to the penalty function $\wp(\theta) = \alpha\|\theta\|_1 + (1-\alpha)\,\|\theta\|_{\mathsf{F}}^2/2$) as the symmetric $\mathbb{R}^{p\times p}$ matrix such that for $1 \leq i, j \leq p$,

$$(\mathrm{Prox}_\gamma(\theta))_{ij} = \begin{cases} 0 & \text{if } |\theta_{ij}| < \alpha\gamma \\ \frac{\theta_{ij}-\alpha\gamma}{1+(1-\alpha)\gamma} & \text{if } \theta_{ij} \geq \alpha\gamma \\ \frac{\theta_{ij}+\alpha\gamma}{1+(1-\alpha)\gamma} & \text{if } \theta_{ij} \leq -\alpha\gamma\,. \end{cases}$$

We consider the following algorithm.

**Algorithm 2.** [Approximate MM algorithm] Fix a step-size $\gamma > 0$. Pick some initial value $\tau^{(0)} \in \mathcal{T}$, $\theta_1^{(0)}, \theta_2^{(0)} \in \mathcal{M}_p^+$. Repeat for $k = 1, \ldots, K$. Given ($\tau^{(k-1)}$, $\theta_1^{(k-1)}$, $\theta_2^{(k-1)}$), do the following:

(1) Compute
$$\theta_1^{(k)} = \mathrm{Prox}_{\gamma\lambda_{1,\tau^{(k-1)}}}\left(\theta_1^{(k-1)} - \gamma\left(S_1(\tau^{(k-1)}) - (\theta_1^{(k-1)})^{-1}\right)\right),$$
(2) compute
$$\theta_2^{(k)} = \mathrm{Prox}_{\gamma\lambda_{2,\tau^{(k-1)}}}\left(\theta_2^{(k-1)} - \gamma\left(S_2(\tau^{(k-1)}) - (\theta_2^{(k-1)})^{-1}\right)\right),$$
(3) compute
$$\tau^{(k)} \stackrel{\text{def}}{=} \mathsf{Argmin}_{t\in\mathcal{T}}\, \mathcal{H}\left(t|\theta_1^{(k)}, \theta_2^{(k)}\right).$$

□

Note that, if instead of a single proximal gradient update in Step (1)-(2), we do a large number proximal gradient updates (an infinite number for the sake of the argument), we recover exactly Algorithm 1. Hence Algorithm 2 is an approximate version of Algorithm 1.

**Remark 1.**
(1) Notice that one can easily compute $\mathcal{H}(\tau+1|\theta_1,\theta_2)$ from $\mathcal{H}(\tau|\theta_1,\theta_2)$ by a rank-one update in $O(p^2)$ number of operations. Hence the computational cost of Step (3) is $O(Tp^2)$. And the total computational cost of one iteration of Algorithm 2 is $O(p^3 + Tp^2)$.
(2) In practice, and as with any gradient descent algorithm, one needs to exercise some care in choosing the step-size $\gamma$. Clearly, too small values of $\gamma$ lead to slow convergence. However, choosing $\gamma$ too large might cause the algorithm to diverge. Another (related) issue is how to guarantee that the matrices $\theta_1^{(k)}$ and $\theta_2^{(k)}$ maintain positive definiteness throughout the iterations. What we show below is that positive definiteness is automatically guaranteed if the step-size $\gamma$ is taken small enough. A nice trade-off that works well from the software engineering viewpoint is to start with a large value of $\gamma$ and to re-initialize the algorithm with a smaller $\gamma$ if at some point positive definiteness is lost. This issue is discussed more extensively in Atchadé et al. (2015).

As suggested in the remark above, Algorithm 2 raises two basic questions. The first question is whether the algorithm is stable, where here by stability we mean whether the algorithm runs without $\theta_1^{(k-1)}$ or $\theta_2^{(k-1)}$ losing positive definiteness. Indeed we notice that Steps (1 and 2) involve taking the inverse of the matrices $\theta_1^{(k-1)}$, and $\theta_2^{(k-1)}$, but there is no guarantee a priori that these matrices are non-singular. Using results established in Atchadé et al. (2015), we answer this question by showing below that if the step-size $\gamma$ is small enough then the algorithm is actually stable. The second basic question is whether the algorithm converges to the optimal value. We address this question below.

For $j \in \{1,2\}$, we set

$$\underline{\lambda}_j \stackrel{\text{def}}{=} \min_{\tau\in\mathcal{T}} \lambda_{j,\tau}, \quad \bar{\lambda}_j \stackrel{\text{def}}{=} \max_{\tau\in\mathcal{T}} \lambda_{j,\tau}, \quad \mu_j \stackrel{\text{def}}{=} \max_{\tau\in\mathcal{T}} \left[\frac{1}{2}\|S_j(\tau)\|_2 + \alpha p \lambda_{j,\tau}\right],$$

$$\mathsf{b}_j \stackrel{\text{def}}{=} \frac{-\mu_j + \sqrt{\mu_j^2 + 2\bar{\lambda}_j(1-\alpha)\frac{n_0}{T}}}{2(1-\alpha)\bar{\lambda}_j}, \quad \mathsf{B}_j \stackrel{\text{def}}{=} \frac{\mu_j + \sqrt{\mu_j^2 + 2\underline{\lambda}_j(1-\alpha)}}{2(1-\alpha)\underline{\lambda}_j}.$$

**Lemma 2.** *Fix $j \in \{1,2\}$. For all $\tau \in \mathcal{T}$, $\hat{\theta}_{j,\tau} \in \mathcal{M}_p^+(\mathsf{b}_j, +\infty)$. Let $\{(\theta_1^{(k)}, \theta_2^{(k)}),\ k \geq 0\}$ be the output of Algorithm 2. If the step-size $\gamma$ satisfies $\gamma \in (0, \mathsf{b}_j^2]$, and $\theta_j^{(0)} \in \mathcal{M}_p^+(\mathsf{b}_j, \mathsf{B}_j)$, then $\theta_j^{(k)} \in \mathcal{M}_p^+(\mathsf{b}_j, \mathsf{B}_j)$, for all $k \geq 0$.*

*Proof.* We present the proof for $j = 1$, the case $j = 2$ being similar. Note that $\hat{\theta}_{1,\tau}$ is the graphical elastic-net estimate based on data $X^{(1)}, \ldots, X^{(\tau)}$. The fact that $\hat{\theta}_{1,\tau}$ exists (and is unique) and satisfies the spectral bound $\lambda_{\mathsf{min}}(\hat{\theta}_{1,\tau}) \geq \mathsf{b}_1$ then follows from known results on the graphical elastic-net (see for instance Lemma 1 of Atchadé et al. (2015)).

The second part of the lemma is similar to Lemma 2 of Atchadé et al. (2015). The idea is to show that if $\theta_1^{(k)} \in \mathcal{M}_p^+(\mathsf{b}_1, \mathsf{B}_1)$ then $\theta_1^{(k+1)} \in \mathcal{M}_p^+(\mathsf{b}_1, \mathsf{B}_1)$. This is proved as follows. Suppose that $\theta_1^{(k)} \in \mathcal{M}_p^+(\mathsf{b}_1, \mathsf{B}_1)$. Hence $\theta_1^{(k)}$ is non-singular. It is well-known (see for instance Parikh and Boyd (2013) Section 4.2) that we can write $\theta_1^{(k+1)}$ as

$$\theta_1^{(k+1)} = \mathsf{Argmin}_{u\in\mathcal{M}_p}\left[\left\langle \nabla g_{1,\tau^{(k)}}(\theta_1^{(k)}), u - \theta_1^{(k)}\right\rangle + \frac{1}{2\gamma}\left\|u - \theta_1^{(k)}\right\|_{\mathsf{F}}^2 + \lambda_{1,\tau^{(k)}}\wp(u)\right].$$

The optimality conditions of this problem implies that there exists $Z \in \mathbb{R}^{p\times p}$, where $Z_{ij} \in [-1,1]$ for all $i,j$ such that

$$\nabla g_{1,\tau^{(k)}}(\theta_1^{(k)}) + \frac{1}{\gamma}\left(\theta_1^{(k+1)} - \theta_1^{(k)}\right) + \lambda_{1,\tau^{(k)}}\left(\alpha Z + (1-\alpha)\theta_1^{(k+1)}\right) = 0.$$

Since $\nabla g_{1,\tau}(\theta) = \frac{\tau}{2T}(S_1(\tau) - \theta^{-1})$, we re-arrange this optimality condition into:

$$\left(1 + (1-\alpha)\lambda_{1,\tau^{(k)}}\gamma\right)\theta_1^{(k+1)} = \theta_1^{(k)} + \frac{\gamma\tau^{(k)}}{2T}\left(\theta_1^{(k)}\right)^{-1} - \gamma\left(\frac{\tau^{(k)}}{2T}S_1(\tau^{(k)}) + \alpha\lambda_{1,\tau^{(k)}}Z\right).$$

Hence, if $\lambda_{\mathsf{min}}(\theta_1^{(k)}) \geq \mathsf{b}_1$, and $\mathsf{b}_1^2 \geq \gamma\tau/(2T)$ (which holds true if $\gamma \leq 2\mathsf{b}_1^2$), and using the fact that $\lambda_{\mathsf{min}}(A+B) \geq \lambda_{\mathsf{min}}(A) + \lambda_{\mathsf{min}}(B)$, we get

$$\lambda_{\mathsf{min}}(\theta_1^{(k+1)}) \geq \frac{1}{1 + (1-\alpha)\bar{\lambda}_1\gamma}\left(\mathsf{b}_1 + \frac{\gamma n_0}{2T}\frac{1}{\mathsf{b}_1} - \gamma\mu_1\right), \tag{5}$$

where $\mu_1 = \max_{\tau\in\mathcal{T}}\left[\frac{1}{2}\|S_1(\tau)\|_2 + \alpha p\lambda_{1,\tau}\right]$, using the fact that $\|Z\|_2 \leq p$. We note that as chosen, $\mathsf{b}_1$ satisfies

$$(1-\alpha)\bar{\lambda}_1\mathsf{b}_1^2 + \mu_1\mathsf{b}_1 - \frac{n_0}{2T} = 0,$$

and this (with some easy algebra) implies that the right hand side of (5) is equal to $\mathsf{b}_1$. Hence $\lambda_{\mathsf{min}}(\theta_1^{(k+1)}) \geq \mathsf{b}_1$. Similarly, if $\lambda_{\mathsf{max}}(\theta_1^{(k)}) \leq \mathsf{B}_1$, then

$$\lambda_{\mathsf{max}}(\theta_1^{(k+1)}) \leq \frac{1}{1+(1-\alpha)\underline{\lambda}_1\gamma}\left(\mathsf{B}_1 + \frac{\gamma}{2}\frac{1}{\mathsf{B}_1} + \gamma\mu_1\right) = \mathsf{B}_1,$$

where the last equality follows from the fact that we have chosen $\mathsf{B}_1$ such that

$$(1-\alpha)\underline{\lambda}_1\mathsf{B}_1^2 - \mu_1\mathsf{B}_1 - \frac{1}{2} = 0.$$

This completes the proof. □

**Remark 3.** The first statement of Lemma 2 implies that the change-point problem (3) has at least one solution. The second part shows that when the step-size $\gamma$ is small enough, all the iterates of the algorithm remains positive definite. We note that the fact that $\alpha < 1$ is crucial in the arguments. The result remains true where $\alpha = 1$, however the arguments is slightly more involved (see Atchadé et al. (2015) Lemma 2). For simplicity we focus in this paper on the case $\alpha \in [0, 1)$.

We now address the issue of convergence. Clearly the function $t \mapsto \mathcal{H}(t|\theta_1, \theta_2)$ is not smooth, nor convex. This implies that Algorithm 2 cannot be analyzed using standard optimization tools. And indeed, we will not be able to establish that the output of Algorithm 2 converges to the minimizer $\hat{\tau}$. Rather, we introduce a containment assumption (Assumption H1) and we show that when it holds, then the output of Algorithm 2 converges to some neighborhood of the true change-point (the existence of this true change-point is part of the assumption).

**H 1.** *There exist $\epsilon > 0$, $c \geq 0$, $\kappa \in [0, 1)$, and $\tau_\star \in \mathcal{T}$ such that the following holds. For any $\tau \in \mathcal{T}$, and for any $\theta_1, \theta_2 \in \mathcal{M}_p^+$ such that $\left\|\theta_1 - \hat{\theta}_{1,\tau}\right\|_F + \left\|\theta_2 - \hat{\theta}_{2,\tau}\right\|_F \leq \epsilon$ we have*

$$|\mathsf{Argmin}_{\,t\in\mathcal{T}}\mathcal{H}(t|\theta_1, \theta_2) - \tau_\star| \leq \kappa|\tau - \tau_\star| + c. \tag{6}$$

**Remark 4.** Plainly, what is imposed in H1 is the existence of a time point $\tau_\star \in \mathcal{T}$ (that we can view as the true change-point), such that anytime we take $\tau \in \mathcal{T}$ that is far from $\tau_\star$ in the sense that $|\tau - \tau_\star| > c/(1-\kappa)$, if $\theta_1, \theta_2$ are sufficiently close to the solutions $\hat{\theta}_{1,\tau}$ and $\hat{\theta}_{2,\tau}$ respectively, then computing $\mathsf{Argmin}_{\,t\in\mathcal{T}}\mathcal{H}(t|\theta_1, \theta_2)$ brings us closer to $\tau_\star$:

$$|\mathsf{Argmin}_{\,t\in\mathcal{T}}\mathcal{H}(t|\theta_1, \theta_2) - \tau_\star| \leq \kappa|\tau - \tau_\star| + c < |\tau - \tau_\star|.$$

This containment assumption is akin to a curvature assumption on the function $t \mapsto \mathcal{H}(t|\theta_1, \theta_2)$ when $\theta_1$ and $\theta_2$ are reasonably close to $\hat{\theta}_{1,\tau}$, $\hat{\theta}_{2,\tau}$, respectively. The assumption seems realistic in settings where the data $X^{(1:T)}$ is indeed drawn from a

Gaussian graphical model with true change-point $\tau_\star$, and parameters $\theta_{\star,1}$, $\theta_{\star,2}$. Indeed in this case, and if $T$ is large enough, for any $\tau$ that is not too close to the boundaries, one expects $\hat{\theta}_{1,\tau}$ and $\hat{\theta}_{2,\tau}$ to be good estimates of $\theta_{\star,1}$ and $\theta_{\star,2}$, respectively. Therefore if $\left\|\theta_1 - \hat{\theta}_{1,\tau}\right\|_{\mathsf{F}} + \left\|\theta_2 - \hat{\theta}_{2,\tau}\right\|_{\mathsf{F}} \leq \epsilon$ for $\epsilon$ small enough, one expect as well $\theta_1$ and $\theta_2$ to be close to $\theta_{\star,1}$ and $\theta_{\star,2}$ respectively. Hence $\mathsf{Argmin}_{t\in\mathcal{T}}\mathcal{H}(t|\theta_1,\theta_2)$ should be close to $\mathsf{Argmin}_{t\in\mathcal{T}}\mathcal{H}(t|\theta_{\star,1},\theta_{\star,2})$, which in turn should be close to $\tau_\star$. Theorem 9 below will make this intuition precise.

□

In the next result we will see that in fact the iterates $\theta_1^{(k)}$ and $\theta_2^{(k)}$ closely track $\theta_{1,\tau^{(k)}}$ and $\theta_{2,\tau^{(k)}}$ respectively. Hence, when H1 holds Equation (6) guarantees that the sequence $\tau^{(k)}$ remains close to $\tau_\star$.

**Theorem 5.** *Suppose that $\gamma \in (0, \mathsf{b}_1^2 \wedge \mathsf{b}_2^2]$, and $\theta_j^{(0)} \in \mathcal{M}_p^+(\mathsf{b}_j, \mathsf{B}_j)$, for $j = 1, 2$. Then*

$$\lim_k \left\|\theta_1^{(k)} - \hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}} = 0, \quad \lim_k \left\|\theta_2^{(k)} - \hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}} = 0.$$

*Furthermore, if H1 holds then*

$$\limsup_{k\to\infty} \left|\tau^{(k)} - \tau_\star\right| \leq \frac{c}{1-\kappa}.$$

*Proof.* See Section 4.1 □

**Remark 6.** Note that the theorem does not guarantee that $\tau^{(k)}$ converges to $\tau_\star$, but rather its conclusion is that for $k$ large $\tau^{(k)}$ stays within $c/(1-\kappa)$ of $\tau_\star$.

We now address the question whether H1 is a realistic assumption. More precisely we will show that the argument highlighted in Remark 4 holds true under some regularity conditions. Suppose that $X^{(1:T)} \stackrel{\text{def}}{=} (X^{(1)}, \ldots, X^{(T)})$ are $p$-dimensional independent random variables such that

$$X^{(1)}, \ldots, X^{(\tau_\star)} \overset{i.i.d.}{\sim} \mathbf{N}(0, \theta_{\star,1}^{-1}), \quad \text{and} \quad X^{(\tau_\star+1)}, \ldots, X^{(T)} \overset{i.i.d.}{\sim} \mathbf{N}(0, \theta_{\star,2}^{-1}), \tag{7}$$

for some unknown change-point $\tau_\star$, and unknown symmetric positive definite precision matrices $\theta_{\star,1} \neq \theta_{\star,2}$. We set $\Sigma_{\star,j} \stackrel{\text{def}}{=} \theta_{\star,j}^{-1}$, and we let $s_j$ denote the number of non-zero entries of $\theta_{\star,j}$, $j = 1, 2$. For an integer $\iota \in \{1, \ldots, p\}$, we define the $\iota$-th restricted eigenvalues of $\Sigma_{\star,j}$ as

$$\underline{\kappa}_j(\iota) \stackrel{\text{def}}{=} \inf\left\{u'(\Sigma_{\star,j})u,\ \|u\|_2 = 1,\ \|u\|_0 \leq \iota\right\},$$
$$\bar{\kappa}_j(\iota) \stackrel{\text{def}}{=} \sup\left\{u'(\Sigma_{\star,j})u,\ \|u\|_2 = 1,\ \|u\|_0 \leq \iota\right\}.$$

We set $s \stackrel{\text{def}}{=} \max(s_1, s_2)$, $\bar{\kappa} \stackrel{\text{def}}{=} \max(\bar{\kappa}_1(2), \bar{\kappa}_2(2))$, $\underline{\kappa} \stackrel{\text{def}}{=} \min(\underline{\kappa}_1(2), \underline{\kappa}_2(2))$, and we set the regularization parameter $\lambda_{j,\tau}$ as

$$\lambda_{1,\tau} \stackrel{\text{def}}{=} \frac{\bar{\kappa}}{\alpha T}\sqrt{48\tau \log(pT)}, \quad \lambda_{2,\tau} \stackrel{\text{def}}{=} \frac{\bar{\kappa}}{\alpha T}\sqrt{48(T-\tau)\log(pT)}, \quad \tau \in \mathcal{T}. \tag{8}$$

We need to assume that the parameter $\alpha \in [0, 1)$ in the regularization term is large enough to produce approximately sparse solutions in (2). To that end, we assume that

$$\frac{\alpha}{1-\alpha} \geq \max\left(\|\theta_{\star,1}\|_\infty, \|\theta_{\star,2}\|_\infty\right). \tag{9}$$

Finally, we assume that the search domain $\mathcal{T}$ is such that for all $\tau \in \mathcal{T}$,

$$\min(\tau, T-\tau) \geq A_1^2 \log(pT), \tag{10}$$

where

$$A_1 \stackrel{\text{def}}{=} \max\left(2\left(\frac{\bar{\kappa}}{\underline{\kappa}}\right)^2, (1280)s^{1/2}\bar{\kappa}(\|\theta_{\star,1}\|_2 \vee \|\theta_{\star,2}\|_2)\right),$$

and

$$\bar{\kappa}\sqrt{\tau \log(pT)} \geq \frac{1}{2\sqrt{3}}(\tau - \tau_\star)_+ \|\theta_{\star,2}^{-1} - \theta_{\star,1}^{-1}\|_\infty,$$
$$\text{and} \quad \bar{\kappa}\sqrt{(T-\tau)\log(pT)} \geq \frac{1}{2\sqrt{3}}(\tau_\star - \tau)_+ \|\theta_{\star,2}^{-1} - \theta_{\star,1}^{-1}\|_\infty, \tag{11}$$

where $x_+ \stackrel{\text{def}}{=} \max(x, 0)$.

**Remark 7.** Assumption (10) is a minimum sample size requirement. See for instance Ravikumar et al. (2011) Theorem 1, and 2 for similar conditions in standard Gaussian graphical model estimation. Here we require to have $\mathcal{T}$ such that $\min(\tau, T-\tau) = O(s\log(pT))$ for all $\tau \in \mathcal{T}$. This obviously implies that we need $T$ to be at least $O(s\log(p))$. It is unclear whether the large constant 1280 in (10) is tight or simply an artifact of our proof techniques.

To understand Assumption (11), note that for $\tau > \tau_\star$, the estimator $\hat{\theta}_{1,\tau}$ in (2) is based on misspecified data $X^{(\tau_\star+1)}, \ldots, X^{(\tau)}$. Hence if $\tau > \tau_\star$ is too far away from $\tau_\star$, the estimators $\hat{\theta}_{1,\tau}$ may behave poorly, particularly if $\theta_{\star,1}$ are $\theta_{\star,2}$ are very different. Assumption (11) rules out such settings, by requiring the search domains $\mathcal{T}$ to be roughly a $\sqrt{T}$ neighborhood of $\tau_\star$. Indeed, suppose that $\tau_\star = \rho_\star T$, for some $\rho_\star \in (0, 1)$. Then it can be easily checked that any search domain of the form $(\tau_\star - r_1 T^{1/2}, \tau_\star + r_2 T^{1/2})$, satisfies (10) and (11) for $T$ large enough, provided that

$$0 < r_1 \leq \frac{2\sqrt{3}\bar{\kappa}\sqrt{\rho_\star \log(pT)}}{\|\theta_{\star,2}^{-1} - \theta_{\star,1}^{-1}\|_\infty}, \quad \text{and} \quad 0 < r_2 \leq \frac{2\sqrt{3}\bar{\kappa}\sqrt{(1-\rho_\star)\log(pT)}}{\|\theta_{\star,2}^{-1} - \theta_{\star,1}^{-1}\|_\infty}.$$

Of course, this search domain is difficult to use in practice since it depends on $\tau_\star$. In practice, we have found that taking $\mathcal{T}$ of the form $(rT, (1-r)T)$ for $r \leq 0.1$ works well, even though it is much wider than what is prescribed by our theory.

□

For $\tau \in \mathcal{T}$, let

$$r_{1,\tau} \stackrel{\text{def}}{=} A_2 \bar{\kappa} \|\theta_{\star,1}\|_2^2 \sqrt{\frac{s_1 \log(pT)}{\tau}}, \quad r_{2,\tau} \stackrel{\text{def}}{=} A_2 \bar{\kappa} \|\theta_{\star,2}\|_2^2 \sqrt{\frac{s_2 \log(pT)}{T-\tau}},$$

where $A_2$ is an absolute constant that can be taken as $16 \times 20 \times \sqrt{48}$. We set $b \stackrel{\text{def}}{=} \min(\lambda_{\mathsf{min}}(\theta_{\star,1}), \lambda_{\mathsf{min}}(\theta_{\star,2}))$, and $B \stackrel{\text{def}}{=} \max(\lambda_{\mathsf{max}}(\theta_{\star,1}), \lambda_{\mathsf{max}}(\theta_{\star,2}))$. We assume that for $j = 1, 2$, and for $\tau \in \mathcal{T}$,

$$r_{j,\tau} \leq \min\left( \frac{\lambda_{\mathsf{min}}(\theta_{\star,j})}{4}, \frac{\|\theta_{\star,j}\|_\infty}{2}, \frac{\|\theta_{\star,j}\|_1}{1 + 8 s_j^{1/2}} \right), \quad r_{j,\tau} \leq \frac{\|\theta_{\star,2} - \theta_{\star,1}\|_{\mathsf{F}}}{2(1 + 8 s^{1/2})}$$

$$\text{and} \quad r_{j,\tau} \leq A_2 \left( \frac{b}{B} \right)^4 \frac{\|\theta_{\star,j}\|_1}{s_j^{1/2}}. \quad (12)$$

**Remark 8.** Condition (12) is mostly technical. As we will see below in Lemma 16, the term $r_{j,\tau}$ is the convergence rate toward $\theta_{\star,j}$ of the estimator $\hat{\theta}_{j,\tau}$, and is expected to converge to 0 with $p, T$ (which implies that the sample size $T$ cannot be too small compared to $\|\theta_{\star,j}\|_2^4 s_j \log(pT)$). Hence according to (12) the matrices $\theta_{\star,1}$ and $\theta_{\star,2}$ need to be such that the terms on the right-hand sides do no vanish faster than the rate $r_{j,\tau}$. In particular $\theta_{\star,1}$ and $\theta_{\star,2}$ should be well-conditioned so that $\lambda_{\mathsf{min}}(\theta_{\star,j})$ and the ratio $b/B$ do not decay too fast.

**Theorem 9.** *Consider the output $\{(\theta_1^{(k)}, \theta_2^{(k)}),\ k \geq 0\}$ of Algorithm 2. Suppose that $\gamma \in (0, \mathsf{b}_1^2 \wedge \mathsf{b}_2^2]$, and $\theta_j^{(0)} \in \mathcal{M}_p^+(\mathsf{b}_j, \mathsf{B}_j)$, for $j = 1, 2$. Suppose that the statistical model underlying the data $X^{(1:T)}$ is as in (7), and that (8)-(12) hold. Suppose also that*

$$\|\theta_{\star,2} - \theta_{\star,1}\|_{\mathsf{F}} \geq 8 A_2 \max\left[ \left( \frac{\lambda_{\mathsf{min}}(\theta_{\star,1})}{\lambda_{\mathsf{max}}(\theta_{\star,1})} \right)^2 \frac{\|\theta_{\star,1}\|_1}{s_1^{1/2}}, \left( \frac{\lambda_{\mathsf{min}}(\theta_{\star,2})}{\lambda_{\mathsf{max}}(\theta_{\star,2})} \right)^2 \frac{\|\theta_{\star,2}\|_1}{s_2^{1/2}} \right]. \quad (13)$$

*Then with probability at least $1 - \frac{8}{pT} - \frac{4}{p^2(1-e^{-C_0})}$, H1 holds with $\epsilon = (1/\sqrt{p}) \min_{\tau \in \mathcal{T}}(r_{1,\tau} \wedge r_{2,\tau})$, $\kappa = 0$, and $c = 4\log(p)/C_0$, where*

$$C_0 \stackrel{\text{def}}{=} \min\left[ \frac{\|\theta_{\star,2} - \theta_{\star,1}\|_{\mathsf{F}}^4}{128 B^4 \|\theta_{\star,2} - \theta_{\star,1}\|_1^2}, \left( \frac{\underline{\kappa}}{\bar{\kappa}} \right)^4 \right].$$

*In particular, we have*

$$\limsup_{k \to \infty} \left| \tau^{(k)} - \tau_\star \right| \leq \frac{4}{C_0} \log(p), \quad (14)$$

*Proof.* See Section 4.2. □

**Remark 10.** The main point of the theorem is that under the assumptions and data generation mechanism described above, the containment assumption H1 holds with probability as least $1-\frac{8}{pT}-\frac{4}{p^2\left(1-e^{-C_0}\right)}$, and where $\epsilon$ can be taken as $\min_\tau r_{1,\tau}\wedge r_{2,\tau}/\sqrt{p}$, $\kappa = 0$, and $c = 4\log(p)/C_0$. Conclusion (14) is then simply a consequence of Theorem 5.

**Remark 11.** We note that the estimation bound in (14) grows with $p$. In classical change-point problems where $p$ is fixed, and $T \to \infty$, it is known (see e.g. Bai (1997)) that with a fixed-magnitude change, the best one can achieve in estimating $\tau$ is $O(1)$. The rate in Theorem 9 suggests that in the high-dimensional setting where $p$ grows the estimation rate for $\tau$ if of order $O(\log(p))$ (see also Roy et al. (2017)). We believe that it is not possible to remove the additional $\log(p)$ factor, although to the best of our knowledge this question is still open. Note that it is customary in the change-point literature to take a re-scaled viewpoint and to define the change point as $a_\star \in (0,1)$ such that $\tau_\star = a_\star T$. In that setting the estimation rate for $a_\star$ is $O(1/T)$ in the classical fixed-dimensional fixed-magnitude change setting, and $O(\log(p)/T)$ in our setting.

2.1. **A stochastic version.** When $T$ is much larger than $p$, Step 3 of Algorithm 2 becomes costly. In such cases, one can gain in efficiency by replacing Step 3 by a Monte Carlo approximation. We explore the use of simulated annealing to approximately solve Step 3 of Algorithm 2. Given $\theta_1, \theta_2 \in \mathcal{M}_p$, and $\beta > 0$, let $\pi_{\beta,\theta_1,\theta_2}$ denote the probability distribution on $\mathcal{T}$ defined as

$$\pi_{\beta,\theta_1,\theta_2}(\tau) = \frac{1}{Z_{\beta,\theta_1,\theta_2}} \exp\left(-\frac{\mathcal{H}(\tau|\theta_1,\theta_2)}{\beta}\right), \quad \tau \in \mathcal{T}.$$

Here, $Z_{\beta,\theta_1,\theta_2}$ is the normalizing constant, and $\beta > 0$ is the cooling parameter, that we shall drive down to zero with the iteration to increase the accuracy of the Monte Carlo approximation. Direct sampling from $\pi_{\beta,\theta_1,\theta_2}$ is typically possible, but this has the same computational cost as Step 3 of Algorithm 2. We will use a Markov Chain Monte Carlo approach which will allow us to make only a small number of calls of the function $\mathcal{H}$, per iteration. Let $\mathcal{K}_{\beta,\theta_1,\theta_2}$ denote a Markov kernel on $\mathcal{T}$ with invariant distribution $\pi_{\beta,\theta_1,\theta_2}$. Typically we will choose $\mathcal{K}_{\beta,\theta_1,\theta_2}$ as a Metropolis-Hastings Markov kernel (we give examples below).

We consider the following algorithm. As in Algorithm 2, $\gamma$ is a given step-size. We choose a decrease sequence of temperature $\beta^{(k)}$ that we use along the iterations.

**Algorithm 3.** Fix a step-size $\gamma > 0$, and a cooling sequence $\{\beta^{(k)}\}$. Pick some initial value $\tau^{(0)} \in \mathcal{T}$, $\theta_1^{(0)}, \theta_2^{(0)} \in \mathcal{M}_p^+$. Repeat for $k = 1, \ldots, K$. Given ($\tau^{(k-1)}$, $\theta_1^{(k-1)}$, $\theta_2^{(k-1)}$), do the following:

(1) Compute

$$\theta_1^{(k)} = \mathrm{Prox}_{\gamma\lambda_{1,\tau^{(k-1)}}}\left(\theta_1^{(k-1)} - \gamma\left(S_1(\tau^{(k-1)}) - (\theta_1^{(k-1)})^{-1}\right)\right),$$

(2) compute

$$\theta_2^{(k)} = \mathrm{Prox}_{\gamma\lambda_{2,\tau^{(k-1)}}}\left(\theta_2^{(k-1)} - \gamma\left(S_2(\tau^{(k-1)}) - (\theta_2^{(k-1)})^{-1}\right)\right),$$

(3) draw

$$\tau^{(k)} \sim \mathcal{K}_{\beta^{(k)},\theta_1^{(k)},\theta_2^{(k)}}(\tau^{(k-1)}, \cdot).$$

□

For most commonly used MCMC kernels, each iteration of Algorithm 3 has a computational cost of $O(p^3)$, which is better than $O(p^3 + Tp^2)$ needed by Algorithm 2, when $T \geq p$. However Algorithm 3 travels along the change-point space $\mathcal{T}$ more slowly. Hence overall, a larger number of iterations would typically be needed for Algorithm 3 to converge. Even after accounting for this slow convergence, Algorithm 3 is still substantially faster than Algorithm 2, as shown in Table 1 and 2. A rigorous analysis of the convergence of Algorithm 3 is beyond the scope of this work, and it left as a possible future research.

2.2. **Extension to multiple change-points.** We extend the method to multiple change-points by binary segmentation. Binary segmentation is a standard method for detecting multiple change-points. The method proceeds by first searching for a single change-point. When a change-point is found the data is split into the two parts defined by the detected change-point. A similar search is then performed on each segment which can result in further splits. This recursive procedure continues until a certain stopping criterion is satisfied. Here we stop the recursion if

$$\ell_\tau + Cp \geq \ell_F,$$

where $\ell_\tau$ is the penalized negative log-likelihood obtained with the additional change-point $\tau$, and $\ell_F$ is the penalized negative log-likelihood without the change-point. The term $Cp$ is a penalty term for model complexity, where $C$ is a user-defined regularization parameter that controls the sparsity of the change-point model (the number of change-points). To the best of our knowledge there is no easy and principled approach for choosing $C$. We identify this as an important issue where more research is needed. Since $C$ controls the number of change-points, in practice one

ad-hoc approach is to set $C$ such that the number of detected change-points is reasonable. This is the approach that we use in the real data analysis. Here we rely on simulation. We explore various scenarios by simulation and found that values of $C$ between $(0, 4)$ produce the best results in our setting.

The binary segmentation algorithm can be defined more precisely as follows. Let us call $\mathcal{J}(X, t_0, t_1)$ the (single) change-point output either by Algorithm 3 or Algorithm 4 when applied to dataset $X$ using sample $X_{t_0}, \ldots, X_{t_1}$, for some $t_0, t_1 \in \mathcal{T}$, $t_0 < t_1$. Let $\mathcal{L}(X, t_0, t_1)$ denote the (penalized) minimum negative log-likelihood achieved on data $X_{t_0}, \ldots, X_{t_1}$. That is,

$$\mathcal{L}(X, t_0, t_1) = \min_{\theta \succ 0} \left[ -\log\det(\theta) + \mathsf{Tr}\left( \theta \left( \frac{1}{t_1 - t_0 + 1} \sum_{t=t_0}^{t_1} X^{(t)} X^{(t)'} \right) \right) + \lambda_{\wp}(\theta) \right].$$

Then the binary-segmentation algorithm $\mathcal{B}(X, t_0, t_1)$ can be written recursively as follows:

**Algorithm 4.** Binary Segmentation

**function** $\mathcal{B}(X, t_0, t_1)$
  $\tau = \mathcal{J}(X, t_0, t_1)$ (apply either algorithm 3 or 4 to data $X_{t_0}, \ldots, X_{t_1}$)
  $\ell_\tau = \mathcal{L}(X, t_0, \tau) + \mathcal{L}(X, \tau + 1, t_1)$
  $\ell_F = \mathcal{L}(X, t_0, t_1)$
  **if** $\ell_\tau + Cp \geq \ell_F$ **then**
    **return** $Null$
  **else**
    **return** $\{\tau, \mathcal{B}(X, t_0, \tau), \mathcal{B}(X, \tau + 1, t_1)\}$
  **end if**
**end function**

We end this section with some words of caution. Binary segmentation is well-known to be a sub-optimal procedure and can perform poorly in some settings (see for instance Fryzlewicz (2014)). The issue is that at each step, binary segmentation is actually fitting a possibly misspecified model – one with a single change-point – to data with possibly multiple change-points. One approach is overcoming this limitation is to extend our proposed algorithms so as to handle directly multiple change-points. We leave this as an important future work.

## 3. Numerical experiments

We investigate the different algorithms presented here in a variety of settings. For all the algorithms investigated the choice of the step-size $\gamma$ and the regularizing parameter $\lambda$ are important. For all experiments, and as suggested by (8), we found that setting $\lambda_{1,\tau} = \lambda\sqrt{\frac{\log\{p\}}{\tau}}$ and $\lambda_{2,\tau} = \lambda\sqrt{\frac{\log\{p\}}{T-\tau}}$ worked well. For the time-comparison in Section 3.1 we used $\lambda = 0.1$ and $\gamma = 3.5$ when $T = 1000$, and we used $\lambda = 0.01$ and $\gamma = 3.5$ when $T = 500$. For the remainder of the experiments we set $\lambda = 0.13$ and $\gamma = 0.25$. For all the experiments the search domain $\mathcal{T}$ is taken as $\{n_0, \ldots, T - n_0\}$, for a minimum sample size $n_0$ from $\{0.01T, 0.05T, 0.1T\}$.

We initialize $\tau^{(0)}$ to a randomly selected value in $\mathcal{T}$. The initial value $\theta_1^{(0)}$ and $\theta_2^{(0)}$ are taken as $\theta_j^{(0)} = (S_j(\tau^{(0)}) + \epsilon I)^{-1}$ where $\epsilon$ is a constant chosen to maintain positive definiteness. For cases where $p < \tau$ and $p < T - \tau$ we used $\epsilon = 0$, while for larger values of $p$ we set $\epsilon = 0.2$.

For the data generation in the simulations, we typically choose $\tau_\star = T/2$ unless otherwise specified, and unless otherwise specified, we generate independently the matrices $\theta_{\star,1}$ and $\theta_{\star,2}$ as follows. First we generate a random symmetric sparse matrix $M$ such that the proportion of non-zero entries is 0.25. We add 4 to all positive entries and subtract 4 from all negative entries. Then we set the actual precision matrix as $\theta_{\star,j} = M + (1 - \lambda_{\min}(M))I_p$ where $\lambda_{\min}(M)$ is the smallest eigenvalue of $M$. The resulting precision matrices contain roughly 25% non-zero off-diagonal elements. For each simulation a new pair of precision matrices was generated as well as the corresponding data set.

For Algorithm 3 we also experimented with a number of MCMC kernel $\mathcal{K}_{\beta,\theta_1,\theta_2}$. We experiment with the independence Metropolis sampler with proposal $\mathbf{U}(n_0, T - n_0)$. We also tried a Random Walk Metropolis with a truncated Gaussian proposal $\mathbf{N}(\tau^{(k-1)}, \sigma^2)$, for some scale parameter $\sigma > 0$. Finally, we also experimented with a mixture of these two Metropolis-Hastings kernels. We found that for our simulations the Independent Metropolis kernel works best, although the mixture kernel also performed well. For the cooling schedule of simulated annealing we use $\beta^{(0)} = 1$, and a geometric decay $\beta^{(n)} = \alpha\beta^{(n-1)}$ with $\alpha = \left(\frac{\beta^{(M)}}{\beta^{(0)}}\right)^{1/M}$ where $\beta^{(M)} = 0.001$, and $M$ is the maximum number of iterations.

An implementation of the algorithms presented here for the Gaussian graphical model context is available in the changepointsHD package, Bybee (2017), available on the Comprehensive R Archive Network (CRAN).

### 3.1. Time comparison.

First we compare the running times of the proposed algorithms and the brute force approach. We consider two settings: $(p = 100, T = 1000)$

| Variant | | Approx. MM | Simulated Annealing |
|---|---|---|---|
| (V1) | Time (Seconds) | 195.95 (48.94) | 3.03 (0.40) |
| | Iterations | 658.68 (82.93) | 662.62 (88.51) |
| (V2) | Time (Seconds) | 0.39 (0.10) | 0.48 (0.46) |
| | Iterations | 1.03 (0.17) | 101.96 (100.29) |

TABLE 1. Run-times of Algorithm 2 and 3 for $(p = 100, T = 1000)$. For comparison the run-time of the brute force algorithm for this problem is 2374.82.

and $(p = 500, T = 500)$. In the setting $(p = 100, T = 1000)$, 100 independent runs of Algorithms 2 and 3 are performed and the average run-times are reported in Table 1. In the setting $(p = 500, T = 500)$ 10 independent runs of Algorithms 2 and 3 are used, and the results are presented in Table 2. We compare these times to results from one simulation run of the brute-force approach, the results of which are given in the description (caption) of Tables 1 and 2.

We consider two stopping criteria for Algorithm 2 or 3. The first criterion stops the iterations if

$$\frac{1}{T}|\tau^{(k)} - \tau_\star| < 0.005 \quad \text{and} \quad \frac{\|\theta_1^{(k)} - \hat{\theta}_1\|_F}{\|\hat{\theta}_1\|_F} + \frac{\|\theta_2^{(k)} - \hat{\theta}_2\|_F}{\|\hat{\theta}_2\|_F} < 0.05, \tag{V1}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are obtained by performing 1000 proximal-gradient steps at the true $\tau$ value. An interesting feature of the proposed approximate MM algorithms is that the change-point sequence $\tau^{(k)}$ can converge well before $\theta_1^{(k)}$ and $\theta_2^{(k)}$. To illustrate this, we also explore the alternative approach of stopping the iterations only based on $\tau^{(k)}$, namely when

$$\frac{1}{T}|\tau^{(k)} - \tau_\star| < 0.005. \tag{V2}$$

Finally, we note that we implement the brute force approach by running 500 proximal-gradient steps for each possible value of $\tau$. Note that 500 iterations is typically smaller than the number of iterations needed to satisfy (V1).

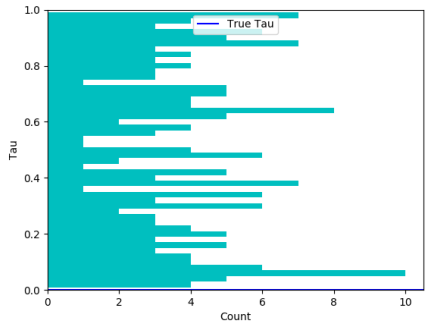
Tables 1 and 2 highlight the benefits of Algorithm 2 and Algorithm 3 as the run-time is several orders of magnitude lower than the brute force approach. Additionally, while Algorithm 3 requires more iterations than Algorithm 2 its run-time is typically smaller. The benefits of Algorithm 3 are particularly clear for large values of $p$ and $T$ (under stopping criterion (V1)). The stopping criteria (V2) highlights the fact that the $\tau^{(k)}$ sequence in the proposed algorithms can converge well before the $\theta$-sequences.

3.2. **Behavior of the algorithm when the change-point is at the edge.** We investigate how the brute force algorithm, Algorithm 2, and Algorithm 3 perform

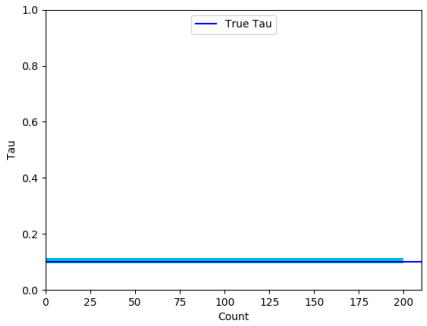
| Variant | | Approx. MM | Simulated Annealing |
|---|---|---|---|
| (V1) | Time (Seconds) | 3554.30 (404.24) | 94.64 (5.50) |
| | Iterations | 939.70 (11.03) | 941.70 (16.23) |
| (V2) | Time (Seconds) | 4.27 (1.10) | 10.96 (8.26) |
| | Iterations | 1.10 (0.32) | 111.20 (90.71) |

TABLE 2. Run-times of Algorithm 2 and 3 for $(p = 500, T = 500)$. For comparison the run-time of the brute force algorithm for this problem is 10854.44.

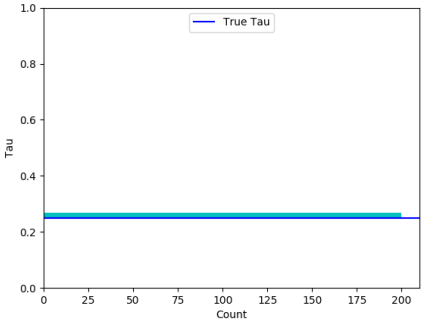
when change-points are non-existent or close to the edges. The results for the brute force algorithm are presented in Figure 1, the results for Algorithm 2 are presented on Figure 2 and the results for Algorithm 3 are presented on Figure 3. For Algorithm 2 and Algorithm 3 the figure contains two subfigures, the first showing the sequences $\{\tau^{(k)}\}$ of solutions produced by the algorithm (trace plots) for all 200 replications, and the second showing a histogram of the final change-point estimate, based on 200 replications. Additionally, a line is included to show the location of the true $\tau$. The trace plots show how quickly each algorithm converges under the various settings. For the brute force algorithm the trace plot is not relevant since the brute force algorithm is not an iterative algorithm. The results suggest that Algorithm 2 and Algorithm 3 have more trouble when the true $\tau$ is close to the edge of the sample. For $\tau = 0.1T$, Algorithm 3 performed slightly better, with 136 simulations ending within 5 units of the true $\tau$ compared to 90 for Algorithm 2.



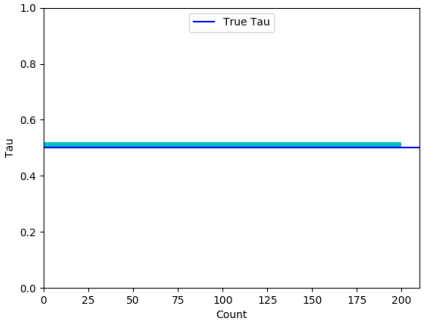

(a) No change-point




(b) Change-point at $\tau = 0.1T$

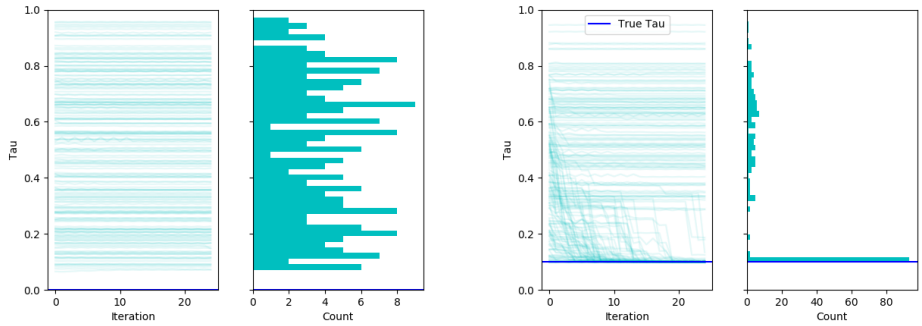



(c) Change-point at $\tau = 0.25T$




(d) Change-point at $\tau = 0.5T$

FIGURE 1. Behavior of the brute force approach as the location of the true change-point is varied. Each plot is a histogram of the change-point estimates based on 200 replications.



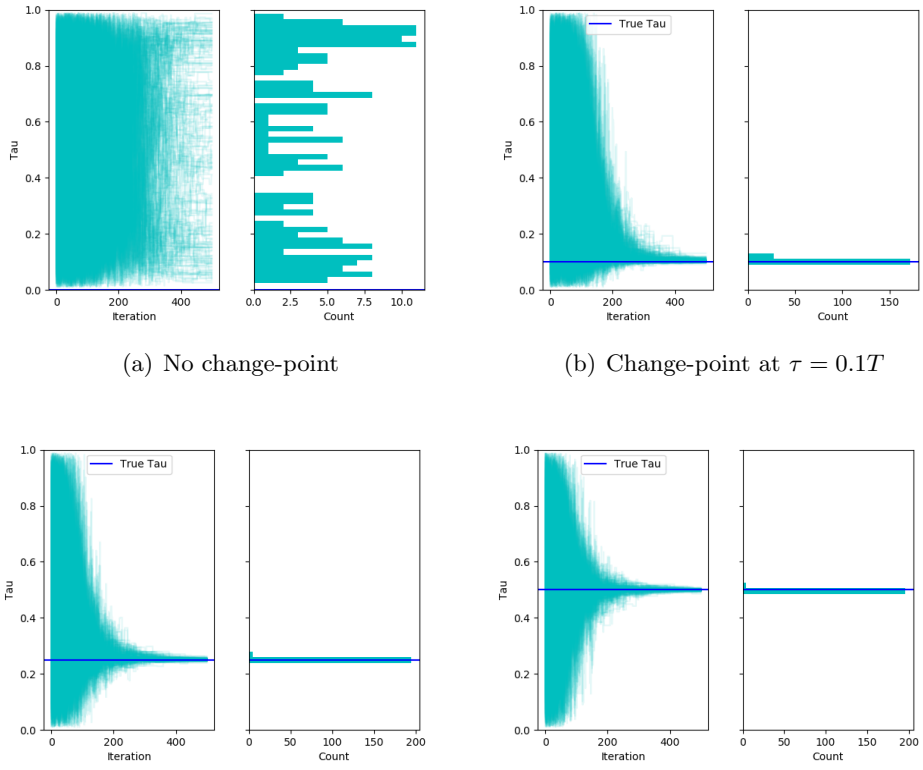

(a) No change-point

(b) Change-point at $\tau = 0.1T$

(c) Change-point at $\tau = 0.25T$

(d) Change-point at $\tau = 0.5T$

FIGURE 2. Behavior of Algorithm 2 as the location of the true change-point is varied. Each plot gives a trace plot of produced estimates, and a histogram of the final change-point estimate. Based on 200 replications.




(a) No change-point

(b) Change-point at $\tau = 0.1T$

(c) Change-point at $\tau = 0.25T$

(d) Change-point at $\tau = 0.5T$

FIGURE 3. Behavior of Algorithm 3 as the location of the true change-point is varied. Each plot gives a trace plot of produced estimates, and a histogram of the final change-point estimate. Based on 200 replications.

3.3. **Behavior of the algorithms when $\theta_1$ and $\theta_2$ are similar.** As $\theta_1$ and $\theta_2$ get increasingly similar, the location of the change-point becomes increasingly more difficult to find. We investigate the behavior of the proposed algorithms in such settings. We generate the true precision matrices $\theta_1$ and $\theta_2$ as follows. We draw a random precision matrix $\theta$ with $q\%$ non-zero off-diagonal elements, and $C_1$ and $C_2$ two random precision matrix with $p\%$ non-zero off-diagonal elements. We choose $C_1$ and $C_2$ to have the same diagonal elements. Then we set $\theta_1 = \theta + C_1$ and $\theta_2 = \theta + C_2$, which are then used to generate the dataset for the experiment. The ratio $p/q$ is a rough indication of the signal. Figure 4-6 show the behavior of the three algorithms for different values of $q$ and $p$. For Algorithms 2 and 3 we found that similar precision matrices sometimes leads the algorithm to converge to the edge of the search domain. This makes sense, since a strong similarity between the two precision matrices implies a weak signal-to-noise ratio, which makes the model with no change-point more attractive. Putting the estimated change-point at the boundary of the search domain is roughly equivalent to fitting a model with no change-point.

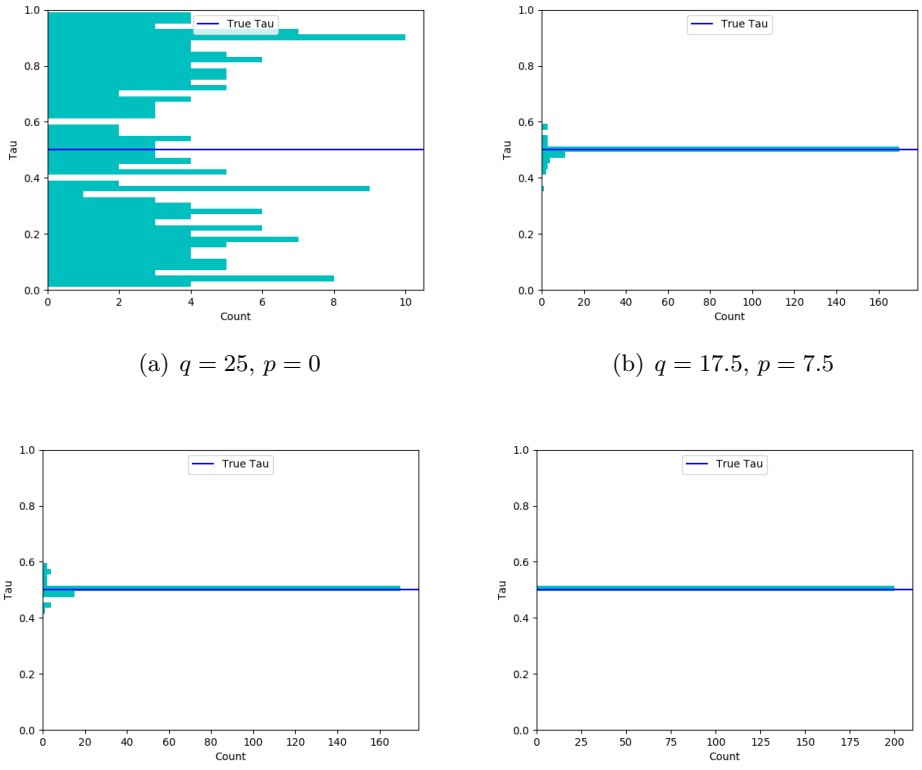
3.4. **Sensitivity to the stopping Criteria in binary segmentation.** This section considers the stopping condition for the binary segmentation algorithm (see Section 2.2) and how it performs with different configurations. A condition is required for determining when the binary segmentation splitting should reject a change-point and stop running. The stopping condition that we use is the following, stop if

$$\ell_\tau + Cp \geq \ell_F,$$

where $\ell_\tau$ is the penalized negative log-likelihood obtained with the additional change-point $\tau$, and $\ell_F$ is the penalized negative log-likelihood without the change-point. The term $C$ is a user-defined parameter.

As mentioned above, the proposed algorithms can diverge when the step-size $\gamma$ is not appropriately selected. In particular the appropriate value of $\gamma$ is highly dependent on the length of the dataset, and the binary segmentation splittings of the data can result in data segments with very different lengths. We use this feature to our advantage. We have chosen not to tune $\gamma$ to the data segment, and to stop the binary segmentation splitting if the sequence $\hat{\theta}_1^{(k)}$ or $\hat{\theta}_2^{(k)}$ appear to diverge. This has the effect of constraining the lengths of the change-point segments from being too small. We achieve this result without directly setting a minimum length constraint – which be hard to do in practice. We found that stopping the algorithm when $||\hat{\theta}_i^{(k)}||_2^2 > 2 \times 10^3$ was sufficient for our data.

In the binary segmentation, since the estimates of $\theta_1$ and $\theta_2$ may not have converged by the end of the search for $\tau$ it may be worth continuing the estimation procedure




(a) $q = 25$, $p = 0$

(b) $q = 17.5$, $p = 7.5$

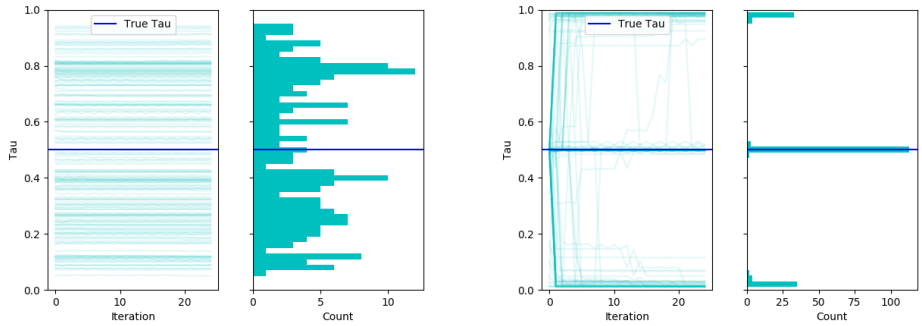
(c) $q = 10$, $p = 15$

(d) $q = 0$, $p = 25$

FIGURE 4. Behavior of the brute force approach for varying signals. Each plot is a histogram of the final change-point estimate. Based on 200 replications.

for $\theta_1$ and $\theta_2$ so that the resulting penalized log-likelihoods are comparable. Hence after each split from the binary segmentation search, we perform an additional 500 iterations to estimate $\theta_1$ and $\theta_2$ at the resulting $\tau$.

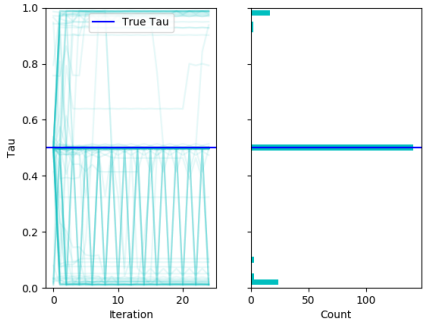
See Figure 7 for a series of heatmaps showing how often the binary segmentation method finds a given number of change-points for different values of $C$. These results suggest that the choice of $C$ in the interval $(0, 4)$ is reasonable. These results are produced using Algorithm 3 for speed, however, the results are identical for the other two algorithms considered. Note that since an additional change-point should always improve the log-likelihood, when $C \leq 0$ we only stop on the secondary stopping condition that $||\hat{\theta}_i^{(k)}||_2^2 > 2 \times 10^3$.



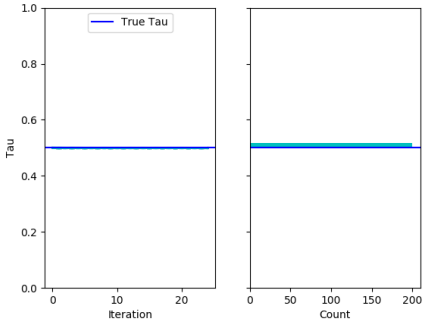

(a) $q = 25$, $p = 0$




(b) $q = 17.5$, $p = 7.5$




(c) $q = 10$, $p = 15$

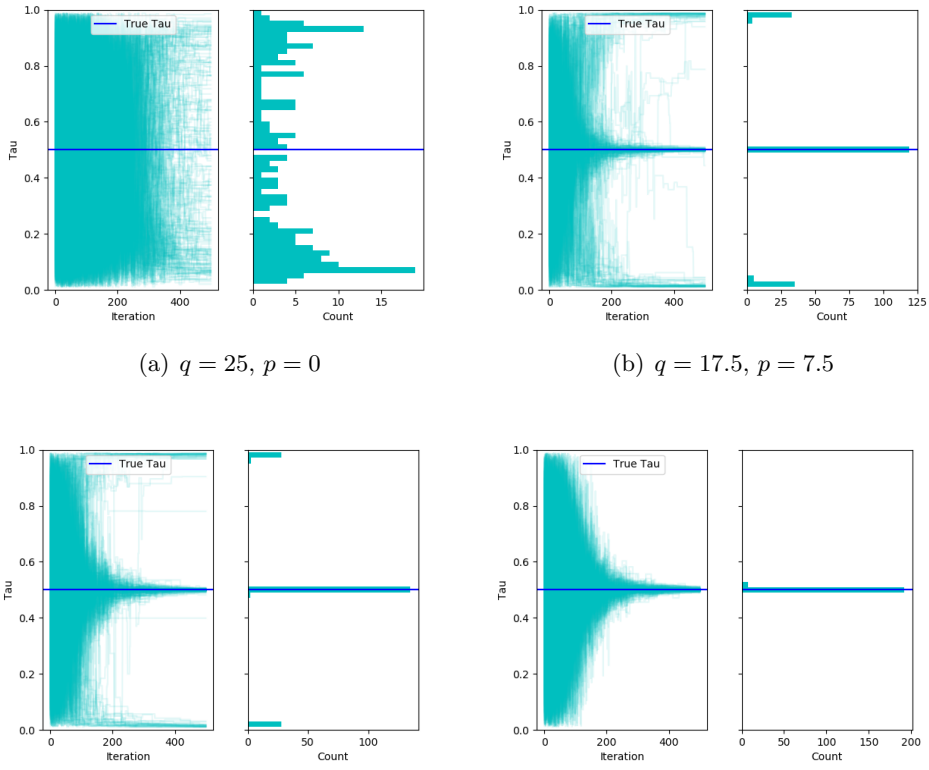



(d) $q = 0$, $p = 25$

FIGURE 5. Behavior of Algorithm 2 for varying signals. Each plot gives a trace plot of produced estimates, and a histogram of the final change-point estimate. Based on 200 replications.

3.5. **High dimensional experiments.** We also investigate the behavior of the proposed algorithms for larger values of $p$. We performed several (100) runs of Algorithm 3 for $T = 1000$, and $p \in \{100, 500, 750, 1000\}$. From these 100 runs we estimate the distributions of the iterates (by boxplots) after $10, 100, 200, \ldots, 1000$ iterations. The results are presented in figure 8. The results show again a very quick convergence toward $\tau_\star$ and this convergence persists even as $p$ gets large.

3.6. **A real data analysis.** In finance and econometrics there is considerable interest in regime-switching models in the context of volatility, particularly because these switches may correspond to real events in the economy (Banerjee and Urga (2005); Beltratti and Morana (2006); Günay (2014); Choi et al. (2010)). However, much of the literature is limited to the low dimensional case, due to the difficulty involved in




(a) $q = 25$, $p = 0$

(b) $q = 17.5$, $p = 7.5$

(c) $q = 10$, $p = 15$

(d) $q = 0$, $p = 25$

FIGURE 6. Behavior of Algorithm 3 for varying signals. Each plot gives a trace plot of produced estimates, and a histogram of the final change-point estimate. Based on 200 replications.

estimating change-points for higher dimensions. We are able to use our method to estimate change-points in the covariance structure of the Standard & Poor's (S&P) 500 – an American stock market index.

Data from the S&P 500 was collected for the period from 2000-01-01 to 2016-03-03. From this initial sample a subset of stocks (or tickers) was selected for which at least 3000 corresponding observations exist. This produced a sample extending from 2004-02-06 to 2016-03-03, consisting of 3039 observations and 436 stocks. We follow a similar data cleaning procedure to Lafferty et al. (2012), who investigate a comparable problem without change-points. For each stock we generate the log returns, $\log \frac{X_t}{X_{t-1}}$, and standarize the resulting returns. Following Lafferty et al. (2012), we then truncate (or clip) all observations beyond three standard deviations of the same mean, thereby

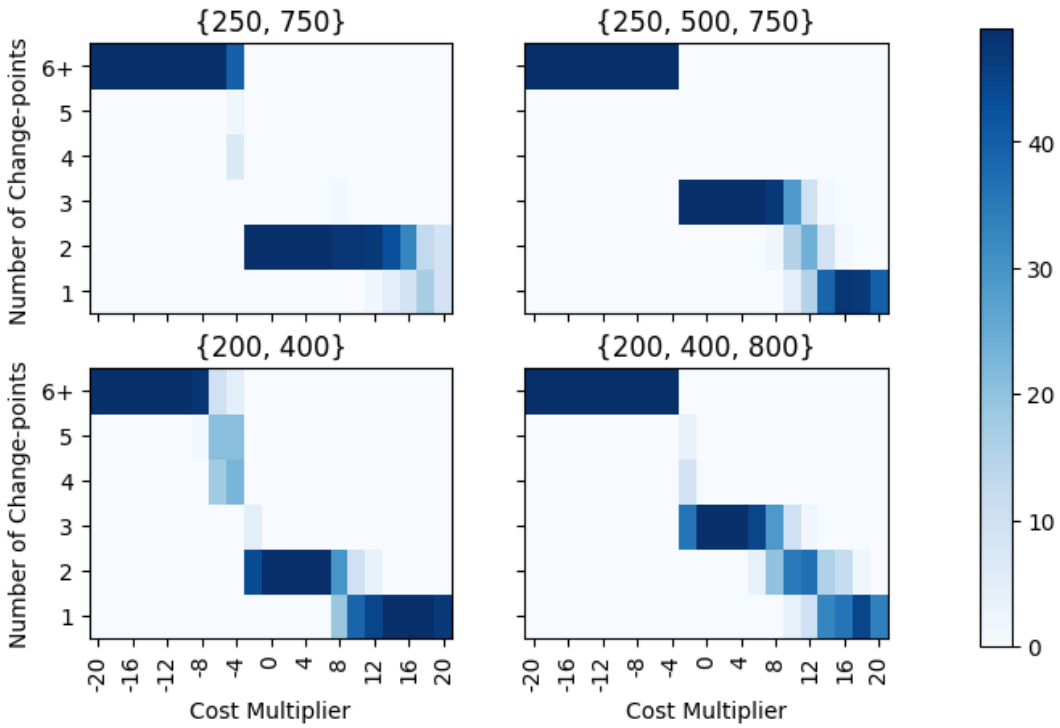



FIGURE 7. Number of change-points detected by binary segmentation as function of the cost multiplier $C$. The set of true change-points is indicated on top of the plots.

limiting unwanted outliers in our sample. The reason for this cleaning procedure is that these outliers often correspond to stock splits instead of meaningful price changes.

For our setting $\lambda = 0.002$ and $\gamma = 0.5$. We initialize $\hat{\theta}^{(0)} = (S(\tau^{(0)}) + I\epsilon)^{-1}$ where $\epsilon = 10^{-4}$ and $\tau^{(0)}$ is selected randomly. After the simulated annealing run the proximal gradient algorithm was run an additional 2000 steps, to produces estimates of $\theta_1$ and $\theta_2$. Here we increase the step-size to $\gamma = 350$ to accelerate the convergence. For the binary segmentation we found that selecting the threshold constant, $C = 0.005$, found a reasonable set of change-points. We found the choice of parameters important in this application, in particular, variation from the values used here can lead the algorithm to diverge. We use the same stopping criterion as with the prior binary-segmentation simulations. That is, a) stop when $\ell_\tau + Cp \geq \ell_F$ or b) stop when $||\hat{\theta}_i^{(k)}||_2^2 > 2 \times 10^3$.

Figure 9 presents the results of the change-point analysis using binary segmentation with Algorithm 4. As a reference we also present the results obtained using binary

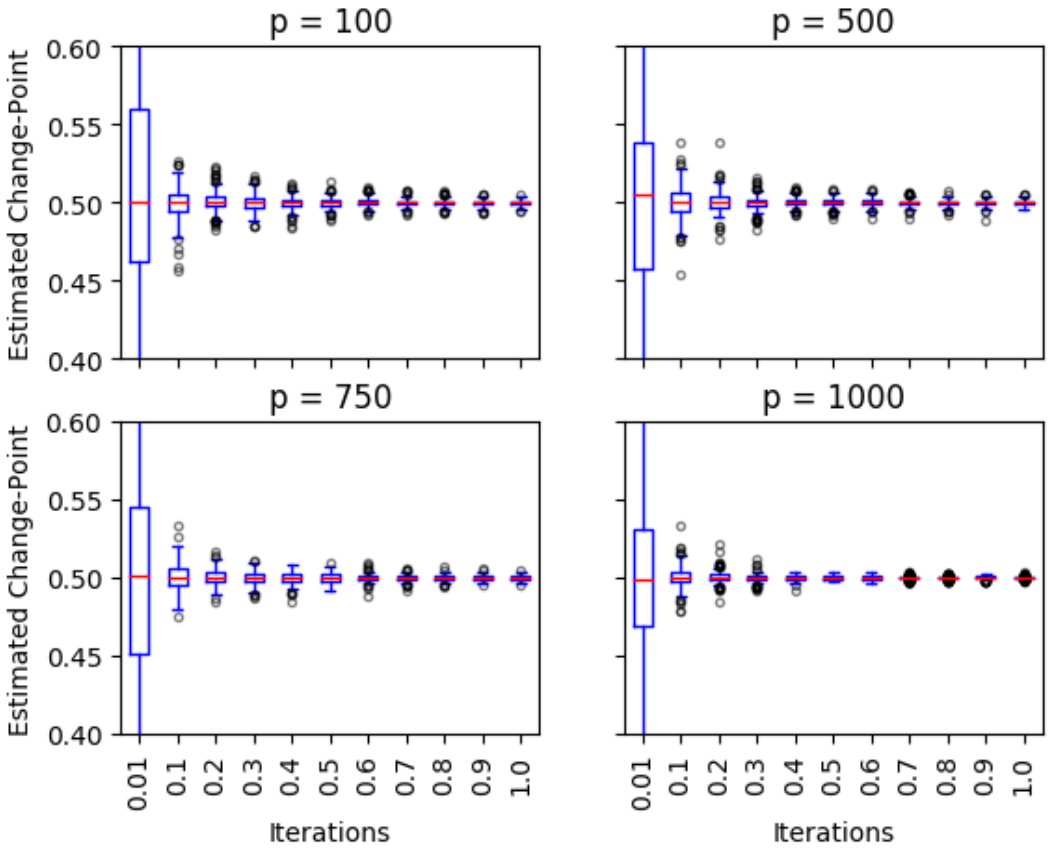


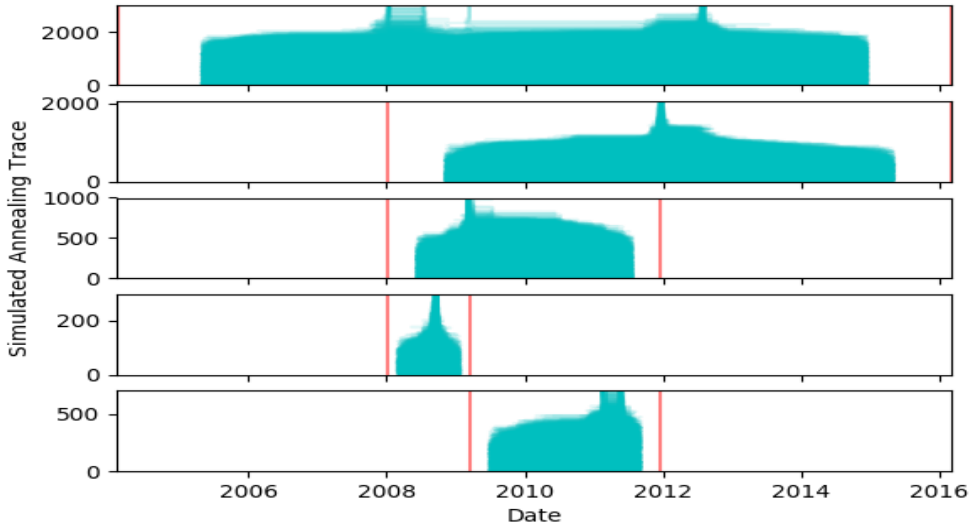

FIGURE 8. Boxplots of the iterates produced by Algorithm 4. Based on 100 replications.

segmentation together with the brute force approach. For the brute force approach, we set $\gamma = 35$ and ran 10 iterations for each possible change-point, before running 2000 steps at $\gamma = 350$ to get the estimates for $\theta_1$ and $\theta_2$. The brute force approach took approximately an hour to run one layer of the search, while simulated annealing took approximately 15 minutes. Figure 9-(a) shows the trace plots from simulated annealing based on 100 replications. The red lines mark the detected time segments. Figure 9-(b) shows the resulting segmentation of the data. We note that simulated annealing and brute force produce slightly different sets of change-points. This brings up an important point: the resulting solution is a local optima. Binary segmentation does introduce an element of path dependency to the results so there may be more than one viable set of change-points – in this particular case, the brute force approach

starts with the first change-point on August 19th 2011 while simulated annealing starts with January 11th 2008.

We next look at how well the estimated change-points correspond to real world events. Our change-point set seems to do a good job of capturing both the Great Recession and a fall in stock prices during August of 2011 related to the European debt crisis and the downgrading of United State's credit-rating. The first change-point in our set is January 11th 2008. The National Bureau of Economic Research (NBER) identifies December of 2007 as the beginning of the Great Recession, which this change-point seems to capture. Additionally, 10 days after the change-point, the Financial Times Stock Exchange (FTSE) would experience its biggest fall since September 11th 2001. The brute force approach places this first change-point earlier in the series on July 23rd 2007, possibly capturing a relatively positive time in the economy before the downturn. The Second change-point occurred on September 15th 2008, the day on which Lehman Brothers filed for bankruptcy protection, one of the key events of the Great Recession (both methods agree on this change-point). The third change-point takes place on March 16th 2009, corresponding to the end of the bear market in the United States. For bthe brute force approach, this change-point is June 2nd 2009 – June of 2009 was when the NBER officially declared the end of the recession. The fourth change-point, on June 1st 2011, and the fifth change-point, on December 21st 2011, likely capture a period of heightened concerns over the possible spread of the European debt crisis to Spain and Italy, during August of 2011. This period also saw the downgrading of the S&P's credit rating of the United States from AAA to AA+. The August 19th 2011 brute force change-point more precisely identifies this August downturn.

Given that the change-point set identified seems sensible, we then investigate what the corresponding $\hat{\theta}$ estimates look like, and whether any interesting conclusions can be drawn from our estimates. Here we focus only on the simulated annealing change-point set. See Figure 10 for a plot of the adjacency matrix for each $\hat{\theta}$ estimate. The black squares correspond to non-zero edges and he yellow boxes correspond to Global Industry Classification Standard (GICS) sectors. These results tell an intuitive story about how the economy behaves during financial crises. Following both the collapse of Lehamn Brother's and the events of August 2011, we see a dramatic increase in connectivity between returns even outside of GICS sectors. To get a better sense of this see Figure 11 for a similar series of plots where edges are summed over each sector. Figure 12 gives an expanded version of the summed edge plot for the first $\hat{\theta}$ estimate, as well as the corresponding sector labels for reference. Again, we can see that during periods of crisis, the off diagonal elements –corresponding to edges




(a) Simulated Annealing trace plots from 100 replications. The red lines represent the prior set of relevant change-points.


(b) Simulated annealing (top) and brute force segmentations of the data.

FIGURE 9. Change-points analysis of the S&P 500 dataset over the period 2004-02-06 to 2016-03-03.

between different sectors – become more significant than during periods of general stability.

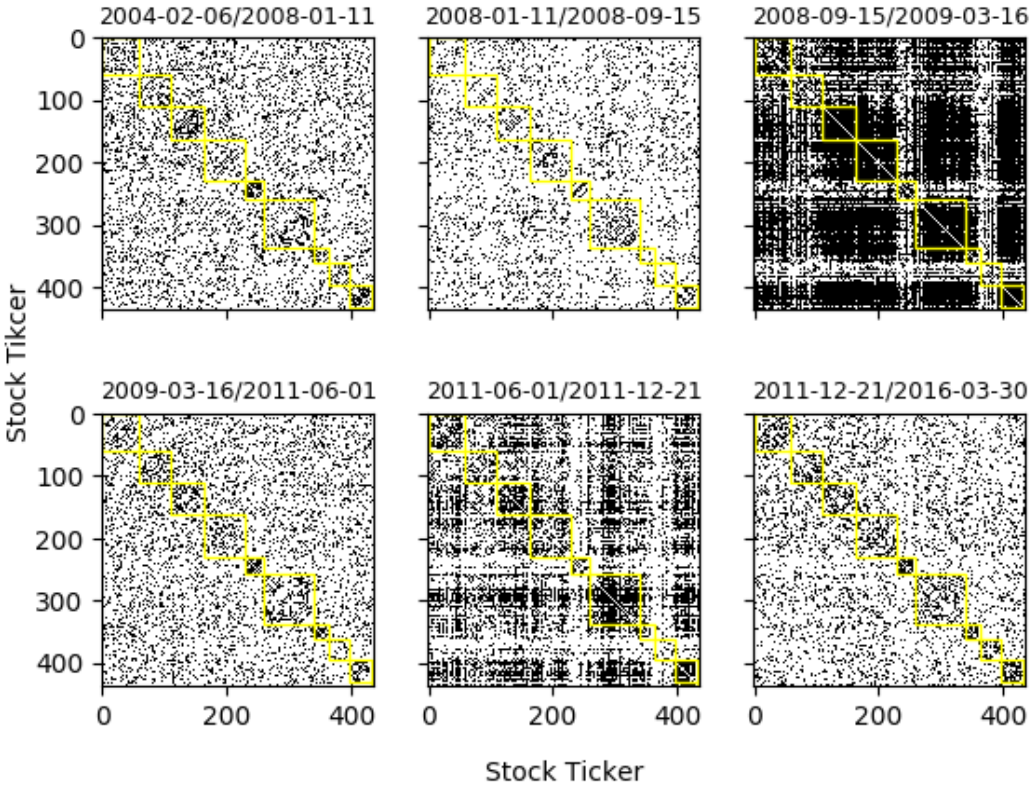



FIGURE 10. Adjacency matrices between stocks based on estimated precision matrices $\hat{\theta}$ for each time segment. A black dot represents an edge between two stocks.

From these figures we can get a sense of which sectors are most affected during times of crisis. To expand upon this some, see Figure 12 for the edge count between each sector and the Financial sector for each $\hat{\theta}$ estimate. We can see that during times of crisis, there is considerable connection between Industrials, Information Technology, Consumer Discretionary, and to a lesser extend Healthcare, and the Financial sector. Consumer Staples, Utilities, and Materials appear to be more stable during these periods and do not experience as much correlation with Financials. This might suggest that our method could be used as a tool to identify investment strategies that are likely to be resilient to periods of crisis in the market.

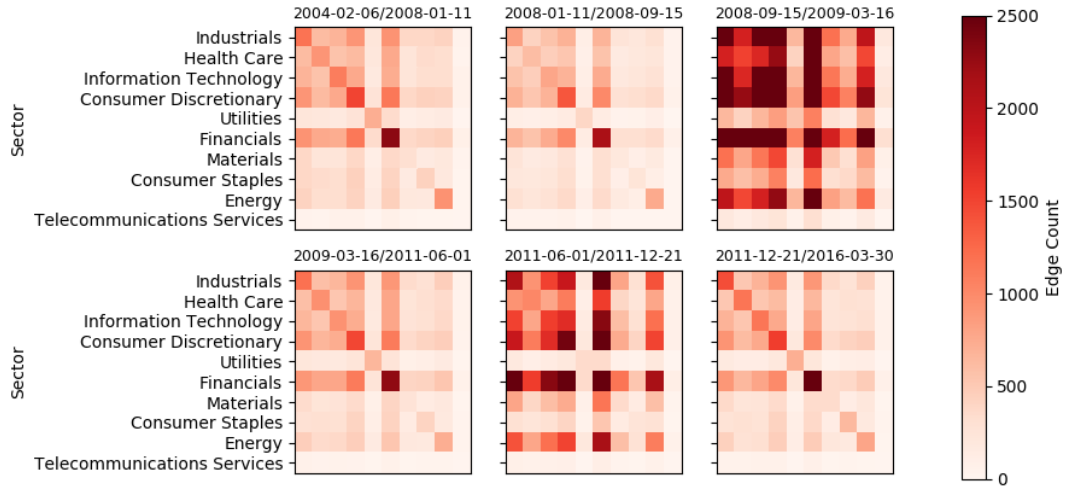



FIGURE 11. Adjacency matrices between sectors for each time segment. Based on the number of edges going from stocks of one sector to another as given by the estimated precision matrices $\hat{\theta}$.

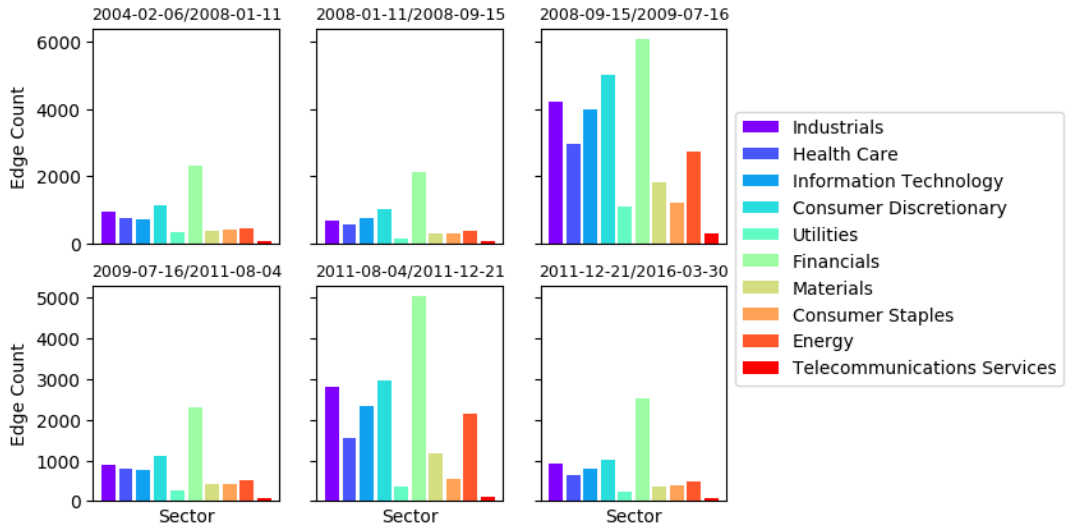



FIGURE 12. Number of edges between the financial sector and the remaining sectors, for each time segment. Based on the estimated precision matrices $\hat{\theta}$.

## 4. PROOFS

4.1. **Proof of Theorem 5.** We will need the following lemma.

**Lemma 12.** *Set*

$$g(\theta) \stackrel{\text{def}}{=} -\log\det(\theta) + \mathsf{Tr}(\theta S),$$

$$\text{and} \quad \phi(\theta) \stackrel{\text{def}}{=} g(\theta) + \lambda\left[\alpha\|\theta\|_1 + \frac{1-\alpha}{2}\|\theta\|_{\mathsf{F}}^2\right], \ \theta \in \mathcal{M}_p^+,$$

*for some symmetric matrix $S$, $\alpha \in (0,1)$, and $\lambda > 0$. Fix $0 < b < B \leq \infty$.*

*(1) For $\theta, \vartheta \in \mathcal{M}_p^+(b,B)$, we have*

$$g(\theta) + \langle \nabla g(\theta), \vartheta - \theta\rangle + \frac{1}{2B^2}\|\vartheta-\theta\|_{\mathsf{F}}^2 \leq g(\vartheta) \leq g(\theta) + \langle \nabla g(\theta), \vartheta - \theta\rangle + \frac{1}{2b^2}\|\vartheta-\theta\|_{\mathsf{F}}^2.$$

*More generally, If $\theta, \vartheta \in \mathcal{M}_p^+$, then*

$$g(\vartheta) - g(\theta) - \langle \nabla g(\theta), \vartheta - \theta\rangle \geq \frac{\|\vartheta-\theta\|_{\mathsf{F}}^2}{4\|\theta\|_2\left(\|\theta\|_2 + \frac{1}{2}\|\vartheta-\theta\|_{\mathsf{F}}\right)}.$$

*(2) Let $\gamma \in (0, b^2]$, and $\theta, \bar{\theta}, \theta_0 \in \mathcal{M}_p^+(b,B)$. Suppose that*

$$\bar{\theta} = \mathrm{Prox}_{\gamma\lambda}\left(\theta - \gamma(S - \theta^{-1})\right),$$

*then*

$$2\gamma\left(\phi(\bar{\theta}) - \phi(\theta_0)\right) + \left\|\bar{\theta} - \theta_0\right\|_{\mathsf{F}}^2 \leq \left(1 - \frac{\gamma}{B^2}\right)\|\theta - \theta_0\|_{\mathsf{F}}^2.$$

*Proof.* The first part of (1) is Lemma 12 of Atchadé et al. (2015), and Part (2) is Lemma 14 of Atchadé et al. (2015). The second part of (1) can be proved along similar lines. For completeness we give the details below.

Take $\theta_0, \theta_1 \in \mathcal{M}_p^+$. By Taylor expansion we have

$$g(\theta_1) - g(\theta_0) - \langle \nabla g(\theta_0), \theta_1 - \theta_0\rangle = -\int_0^1 \left\langle (\theta_0 + tH)^{-1} - \theta_0^{-1}, H\right\rangle \mathrm{d}t,$$

where $H \stackrel{\text{def}}{=} \theta_1 - \theta_0$. We have $(\theta_0 + tH)^{-1} - \theta_0^{-1} = -t\theta_0^{-1}H(\theta_0 + tH)^{-1}$, which leads to

$$g(\theta_1) - g(\theta_0) - \langle \nabla g(\theta_0), \theta_1 - \theta_0\rangle = \int_0^1 \mathsf{Tr}\left(\theta_0^{-1}H(\theta_0 + tH)^{-1}H\right) t\mathrm{d}t.$$

If $\theta_0 = \sum_{i=1}^p \rho_j u_j u_j'$ is the eigendecomposition of $\theta_0$, we see that $\mathsf{Tr}\left(\theta_0^{-1} H(\theta_0 + tH)^{-1} H\right) = \sum_{j=1}^p \frac{1}{\rho_j} u_j' H(\theta_0 + tH)^{-1} H u_j$. Hence

$$g(\theta_1) - g(\theta_0) - \langle \nabla g(\theta_0), \theta_1 - \theta_0 \rangle \geq \sum_{j=1}^p \|Hu_j\|_2^2 \int_0^1 \frac{t\mathrm{d}t}{\|\theta_0\|_2 \left(\|\theta_0\|_2 + t\|H\|_{\mathsf{F}}\right)}$$
$$\geq \frac{\sum_{j=1}^p \|Hu_j\|_2^2}{4\|\theta_0\|_2 \left(\|\theta_0\|_2 + \frac{1}{2}\|H\|_{\mathsf{F}}\right)},$$

and the result follows by noting that $\sum_{j=1}^p \|Hu_j\|_2^2 = \|H\|_{\mathsf{F}}^2$.

□

Set

$$\mathcal{F}(\tau, \theta_1, \theta_2) = g_{1,\tau}(\theta_1) + \lambda_{1,\tau} p(\theta) + g_{2,\tau}(\theta_2) + \lambda_{2,\tau} p(\theta_2),$$

$\underline{\mathcal{F}} = \mathcal{F}(\hat{\tau}, \hat{\theta}_{1,\hat{\tau}}, \hat{\theta}_{1,\hat{\tau}})$ the value of Problem (3), and $\mathcal{F}_k = \mathcal{F}(\tau^{(k)}, \theta_1^{(k)}, \theta_2^{(k)}) - \underline{\mathcal{F}}$.

**Lemma 13.** *Suppose that $\gamma \in (0, \mathsf{b}_1^2 \wedge \mathsf{b}_2^2]$, and for $j = 1, 2$, $\theta_j^{(0)} \in \mathcal{M}_p^+(\mathsf{b}_j, \mathsf{B}_j)$. Then $\lim_k \left\|\theta_1^{(k)} - \hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}} = 0$, $\lim_k \left\|\theta_2^{(k)} - \hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}} = 0$. Furthermore the sequence $\{\mathcal{F}_k\}$ is non-increasing, and $\lim_k \mathcal{F}_k$ exists.*

*Proof.* We know from Lemma 2 that for $\gamma \in (0, \mathsf{b}_1^2 \wedge \mathsf{b}_2^2]$, and $\theta_j^{(0)} \in \mathcal{M}_p^+(\mathsf{b}_j, \mathsf{B}_j)$, we have $\theta_j^{(k)} \in \mathcal{M}_p^+(\mathsf{b}_j, \mathsf{B}_j)$ for all $k \geq 0$, for $j = 1, 2$. We have,

$$\mathcal{F}_{k+1} - \mathcal{F}_k = \mathcal{F}(\tau^{(k+1)}, \theta_1^{(k+1)}, \theta_2^{(k+1)}) - \mathcal{F}(\tau^{(k)}, \theta_1^{(k+1)}, \theta_2^{(k+1)})$$
$$+ \mathcal{F}(\tau^{(k)}, \theta_1^{(k+1)}, \theta_2^{(k+1)}) - \mathcal{F}(\tau^{(k)}, \theta_1^{(k)}, \theta_2^{(k)}).$$

By definition, $\mathcal{F}(\tau^{(k+1)}, \theta_1^{(k+1)}, \theta_2^{(k+1)}) - \mathcal{F}(\tau^{(k)}, \theta_1^{(k+1)}, \theta_2^{(k+1)}) \leq 0$, and by Lemma 12-Part(2),

$$\mathcal{F}(\tau^{(k)}, \theta_1^{(k+1)}, \theta_2^{(k+1)}) - \mathcal{F}(\tau^{(k)}, \theta_1^{(k)}, \theta_2^{(k)})$$
$$\leq -\frac{1}{2\gamma} \left\|\theta_1^{(k+1)} - \theta_1^{(k)}\right\|_{\mathsf{F}}^2 - \frac{1}{2\gamma} \left\|\theta_2^{(k+1)} - \theta_2^{(k)}\right\|_{\mathsf{F}}^2$$

It follows that

$$\mathcal{F}_{k+1} \leq \mathcal{F}_k - \frac{1}{2\gamma} \left\|\theta_1^{(k+1)} - \theta_1^{(k)}\right\|_{\mathsf{F}}^2 - \frac{1}{2\gamma} \left\|\theta_2^{(k+1)} - \theta_2^{(k)}\right\|_{\mathsf{F}}^2,$$

which implies that

$$\lim_k \left\|\theta_1^{(k+1)} - \theta_1^{(k)}\right\|_{\mathsf{F}} = 0, \quad \text{and} \quad \lim_k \left\|\theta_2^{(k+1)} - \theta_2^{(k)}\right\|_{\mathsf{F}} = 0. \tag{15}$$

It also implies that the sequence $\{\mathcal{F}_k\}$ is non-increasing and bounded from below by 0. Hence converges. Another application of Lemma 12 gives

$$\begin{aligned}2\gamma\left(\mathcal{F}(\tau^{(k)},\theta_1^{(k+1)},\theta_2^{(k+1)})-\mathcal{F}(\tau^{(k)},\hat{\theta}_{1,\tau^{(k)}},\hat{\theta}_{2,\tau^{(k)}})\right)&\\ +\left\|\theta_1^{(k+1)}-\hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}}^2+\left\|\theta_2^{(k+1)}-\hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}}^2&\\ \leq\left(1-\frac{\gamma}{\mathsf{B}_1^2}\right)\left\|\theta_1^{(k)}-\hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}}^2+\left(1-\frac{\gamma}{\mathsf{B}_2^2}\right)\left\|\theta_2^{(k)}-\hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}}^2.&\end{aligned}$$

And notice that $\mathcal{F}(\tau^{(k)},\theta_1^{(k+1)},\theta_2^{(k+1)})-\mathcal{F}(\tau^{(k)},\hat{\theta}_{1,\tau^{(k)}},\hat{\theta}_{2,\tau^{(k)}})\geq 0$. Hence

$$\begin{aligned}\left\|\theta_1^{(k+1)}-\hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}}^2+\left\|\theta_2^{(k+1)}-\hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}}^2&\\ \leq\left(1-\frac{\gamma}{\mathsf{B}_1^2}\right)\left\|\theta_1^{(k)}-\hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}}^2+\left(1-\frac{\gamma}{\mathsf{B}_2^2}\right)\left\|\theta_2^{(k)}-\hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}}^2,&\end{aligned}$$

which can be written as

$$\begin{aligned}\frac{\gamma}{\mathsf{B}_1^2}\left\|\theta_1^{(k)}-\hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}}^2+\frac{\gamma}{\mathsf{B}_2^2}\left\|\theta_2^{(k)}-\hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}}^2\leq\left\|\theta_1^{(k+1)}-\theta_1^{(k)}\right\|_{\mathsf{F}}^2+\left\|\theta_2^{(k+1)}-\theta_2^{(k)}\right\|_{\mathsf{F}}^2&\\ -2\left\langle\theta_1^{(k+1)}-\theta_1^{(k)},\theta_1^{(k+1)}-\hat{\theta}_{1,\tau^{(k)}}\right\rangle-2\left\langle\theta_2^{(k+1)}-\theta_2^{(k)},\theta_2^{(k+1)}-\hat{\theta}_{2,\tau^{(k)}}\right\rangle.&\end{aligned}$$

Since $\{\theta_1^{(k)}\}$, $\{\theta_2^{(k)}\}$ $\{\hat{\theta}_{1,\tau^{(k)}}\}$, and $\{\hat{\theta}_{2,\tau^{(k)}}\}$ are bounded sequence, and given (15), letting $k\to\infty$, we conclude that

$$\lim_k\left\|\theta_1^{(k)}-\hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}}=0,\quad\text{and}\quad\lim_k\left\|\theta_2^{(k)}-\hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}}=0.$$

□

*Proof of Theorem 5.* Let $\epsilon>0$ as in H1. By Lemma 13, there exist $k_0\geq 1$ such that for all $k\geq k_0$, $\left\|\theta_1^{(k+1)}-\hat{\theta}_{1,\tau^{(k)}}\right\|_{\mathsf{F}}\leq\epsilon$, and $\left\|\theta_2^{(k+1)}-\hat{\theta}_{2,\tau^{(k)}}\right\|_{\mathsf{F}}\leq\epsilon$. Since

$$\tau^{(k+1)}=\mathsf{Argmin}_{t\in\mathcal{T}}\,\mathcal{H}\left(t|\theta_1^{(k+1)},\theta_2^{(k+1)}\right),$$

using H1 we conclude that for all $k\geq k_0$,

$$\left|\tau^{(k+1)}-\tau_\star\right|\leq\kappa\left|\tau^{(k)}-\tau_\star\right|+c\leq\kappa^{k-k_0+1}\left|\tau^{(k_0)}-\tau_\star\right|+\frac{c}{1-\kappa},$$

which implies the stated result. □

4.2. **Proof of Theorem 9.** We introduce some more notation. Given $M \in \mathbb{R}^{p\times p}$ the sparsity structure of $M$ is the matrix $\delta \in \{0,1\}^{p\times p}$ such that $\delta_{jk} = \mathbf{1}_{\{|M_{jk}|>0\}}$. In particular we will write $\delta_{\star,j}$ $(j = 1, 2)$ to denote the sparsity structure of $\theta_{\star,j}$. Given matrices $A \in \mathbb{R}^{p\times p}$, and $\delta \in \{0,1\}^{p\times p}$, we will use the notation $A_\delta$ (resp. $A_{\delta^c}$) to denote the component-wise product of $A$ and $\delta$ (resp $A$ and $1-\delta$). Given $j \in \{1,2\}$, we define

$$\mathcal{C}_j \stackrel{\text{def}}{=} \left\{ M \in \mathcal{M}_p : \ \|M_{\delta^c_{\star,j}}\|_1 \leq 7\|M_{\delta_{\star,j}}\|_1. \right\}. \tag{16}$$

We will need the following deviation bound.

**Lemma 14.** *Suppose that* $X_i \stackrel{ind}{\sim} \boldsymbol{N}(0, \theta_i^{-1})$, $i = 1, \ldots, N$, *where* $\theta_i \in \mathcal{M}_p^+$. *We set* $\Sigma_i \stackrel{\text{def}}{=} \theta_i^{-1}$, *and define*

$$\underline{\kappa}_i(2) \stackrel{\text{def}}{=} \inf\left\{u'\Sigma_i u, \ \|u\|_2 = 1, \ \|u\|_0 \leq 2\right\}, \ \ \bar{\kappa}_i(2) \stackrel{\text{def}}{=} \sup\left\{u'\Sigma_i u, \ \|u\|_2 = 1, \ \|u\|_0 \leq 2\right\},$$

*and suppose that* $\underline{\kappa}_i(2) > 0$ *for* $i = 1, \ldots, N$. *Set* $G_N \stackrel{\text{def}}{=} N^{-1}\sum_{i=1}^N (X_i X_i' - \theta_i^{-1})$. *Then for* $0 < \delta \leq 2\left(\frac{\min_k \underline{\kappa}_k(2)}{\max_k \bar{\kappa}_k(2)}\right)^2$, *we have*

$$\mathbb{P}\left(\|G_N\|_\infty > \left(\max_k \bar{\kappa}_k(2)\right)\delta\right) \leq 4p^2 e^{-\frac{N\delta^2}{4}}.$$

*Proof.* The proof is similar to the proof of Lemma 1 of Ravikumar et al. (2010), which itself builds on Bickel and Levina (2008). For $1 \leq i, j \leq p$, arbitrary, set $Z_{ij}^{(k)} = X_{k,i}X_{k,j}$, and $\sigma_{ij}^{(k)} = \Sigma_{k,ij}$, so that the $(i,j)$-th component of $G_N$ is $N^{-1}\sum_{k=1}^N (Z_{ij}^{(k)} - \sigma_{ij}^{(k)})$. Suppose that $i \neq j$. The case $i = j$ is simpler. It is easy to check that

$$\sum_{k=1}^N \left[Z_{ij}^{(k)} - \sigma_{ij}^{(k)}\right] = \frac{1}{4}\sum_{k=1}^N \left[(X_{k,i} + X_{k,j})^2 - \sigma_{ii}^{(k)} - \sigma_{jj}^{(k)} - 2\sigma_{ij}^{(k)}\right] - \frac{1}{4}\sum_{k=1}^N \left[(X_{k,i} - X_{k,j})^2 - \sigma_{ii}^{(k)} - \sigma_{jj}^{(k)} + 2\sigma_{ij}^{(k)}\right].$$

Notice that $X_{k,i} + X_{k,j} \sim \mathbf{N}(0, \sigma_{ii}^{(k)} + \sigma_{jj}^{(k)} + 2\sigma_{ij}^{(k)})$, and $X_{k,i} - X_{k,j} \sim \mathbf{N}(0, \sigma_{ii}^{(k)} + \sigma_{jj}^{(k)} - 2\sigma_{ij}^{(k)})$. It follows that for all $x \geq 0$,

$$\mathbb{P}\left[\left|\sum_{k=1}^N \left[Z_{ij}^{(k)} - \sigma_{ij}^{(k)}\right]\right| > x\right] \leq \mathbb{P}\left[\left|\sum_{k=1}^N a_{ij}^{(k)}(W_k - 1)\right| > 2x\right] + \mathbb{P}\left[\left|\sum_{k=1}^N b_{ij}^{(k)}(W_k - 1)\right| > 2x\right],$$

where $W_{1:N} \overset{i.i.d.}{\sim} \chi^2_1$, $a^{(k)}_{ij} = \sigma^{(k)}_{ii} + \sigma^{(k)}_{jj} + 2\sigma^{(k)}_{ij}$, and $b^{(k)}_{ij} = \sigma^{(k)}_{ii} + \sigma^{(k)}_{jj} - 2\sigma^{(k)}_{ij}$. For any $x \geq 0$ and a sequence $a = (a_1, \ldots, a_N)$ of positive numbers, with $|a|_\infty = \max_i |a_i|$, $|a|_2 = \sqrt{\sum_i a_i^2}$, we write

$$2x = 2|a|_2 \left(\frac{x}{2|a|_2}\right) + 2|a|_\infty \left(\frac{4|a|_2^2}{2x|a|_\infty}\right)\left(\frac{x}{2|a|_2}\right)^2.$$

Therefore if $2x|a|_\infty \leq 4|a|_2^2$, we can apply Lemma 1 of Laurent and Massart (2000) to conclude that

$$\mathbb{P}\left(\left|\sum_{k=1}^N a_k(W_k - 1)\right| \geq 2x\right) \leq 2e^{-\frac{x^2}{4|a|_2^2}}.$$

In particular, we can apply the above bound with $x = |a|_\infty N\delta$ for $\delta \in (0, \frac{2\min_j a_i^2}{\max_i a_i^2}]$ to get that

$$\mathbb{P}\left(\left|\sum_{k=1}^N a_k(W_k - 1)\right| \geq 2|a|_\infty N\delta\right) \leq 2e^{-\frac{N\delta^2}{4}}.$$

In the particular case above, $a^{(k)}_{ij} = \sigma^{(k)}_{ii} + \sigma^{(k)}_{jj} + 2\sigma^{(k)}_{ij} = u'\Sigma^{(k)}u$, where $u_i = u_j = 1$, and $u_r = 0$ for $r \notin \{i, j\}$. And

$$\frac{\min_k u'\Sigma^{(k)}u}{\max_k u'\Sigma^{(k)}u} \geq \frac{\min_k \underline{\kappa}_k(2)}{\max_k \bar{\kappa}(2)}.$$

A similar bound holds for $b^{(k)}_{ij}$. The lemma follows from a standard union-sum argument.

□

The following event plays an important role in the analysis.

$$\mathcal{E}_n \overset{\text{def}}{=} \bigcap_{\tau \in \mathcal{T}} \left\{ \frac{1}{\lambda_{1,\tau}} \|\nabla g_{1,\tau}(\theta_{\star,1})\|_\infty \leq \frac{\alpha}{2}, \text{ and } \frac{1}{\lambda_{2,\tau}} \|\nabla g_{2,\tau}(\theta_{\star,2})\|_\infty \leq \frac{\alpha}{2} \right\}, \tag{17}$$

**Lemma 15.** *Under the assumptions of the theorem*

$$\mathbb{P}(\mathcal{E}_n) \geq 1 - \frac{8}{pT}.$$

*Proof.* We have

$$\mathbb{P}(\mathcal{E}_n^c) \leq \mathbb{P}\left(\max_{\tau \in \mathcal{T}} \frac{1}{\lambda_{1,\tau}} \|\nabla g_{1,\tau}(\theta_{\star,1})\|_\infty > \frac{\alpha}{2}\right) + \mathbb{P}\left(\max_{\tau \in \mathcal{T}} \frac{1}{\lambda_{2,\tau}} \|\nabla g_{2,\tau}(\theta_{\star,2})\|_\infty > \frac{\alpha}{2}\right).$$

We show how to bound the first term. A similar bound follows for $g_{2,\tau}$ by working on the reversed sequence $X^{(T)}, \ldots, X^{(1)}$. We have $\nabla g_{1,\tau}(\theta) = \frac{\tau}{2T}(S_1(\tau) - \theta^{-1})$. Setting $U^{(t)} \overset{\text{def}}{=} X^{(t)}(X^{(t)})' - \mathbb{E}\left(X^{(t)}(X^{(t)})'\right)$, we can write

$$\nabla g_{1,\tau}(\theta_{\star,1}) = \frac{1}{2T}\sum_{t=1}^{\tau} U^{(t)} + \frac{(\tau - \tau_\star)_+}{2T}(\theta_{\star,2}^{-1} - \theta_{\star,1}^{-1}),$$

where $a_+ \stackrel{\text{def}}{=} \max(a, 0)$. Hence by a standard union-bound argument,

$$\mathbb{P}\left(\max_{\tau\in\mathcal{T}} \frac{1}{\lambda_{1,\tau}}\|\nabla g_{1,\tau}(\theta_{\star,1})\|_\infty > \frac{\alpha}{2}\right) \leq \sum_{\tau\in\mathcal{T}} \mathbb{P}\left(\left\|\sum_{t=1}^{\tau} U^{(t)}\right\|_\infty > \alpha\lambda_{1,\tau}T - (\tau-\tau_\star)_+\|\theta_{\star,2}^{-1} - \theta_{\star,1}^{-1}\|_\infty\right).$$

Given the choice of $\lambda_{1,\tau}$ in (8), $\alpha\lambda_{1,\tau}T/2 = 2\sqrt{3}\bar{\kappa}\sqrt{\tau\log(pT)} \geq (\tau-\tau_\star)_+\|\theta_{\star,2}^{-1}-\theta_{\star,1}^{-1}\|_\infty$, by assumption (11). In view of (10) we can apply Lemma 14 to deduce that

$$\begin{aligned}\mathbb{P}\left(\max_{\tau\in\mathcal{T}} \frac{1}{\lambda_{1,\tau}}\|\nabla g_{1,\tau}(\theta_{\star,1})\|_\infty > \frac{\alpha}{2}\right) &\leq \sum_{\tau\in\mathcal{T}} \mathbb{P}\left(\left\|\frac{1}{\tau}\sum_{t=1}^{\tau} U^{(t)}\right\|_\infty > \frac{\alpha\lambda_{1,\tau}T}{2\tau}\right)\\ &\leq 4Tp^2 e^{-\frac{\tau}{4}\left(\frac{\alpha\lambda_{1,\tau}T}{2\tau\bar{\kappa}}\right)^2}\\ &\leq 4\exp\left(2\log(pT) - 3\log(pT)\right) \leq \frac{4}{pT}.\end{aligned}$$

□

**Lemma 16.** *Under the assumptions of the theorem, and on the event $\mathcal{E}_n$, we have*

$$\left\|\hat{\theta}_{1,\tau} - \theta_{\star,1}\right\|_{\mathsf{F}} \leq A\bar{\kappa}\|\theta_{\star,1}\|_2^2\sqrt{\frac{s_1\log(pT)}{\tau}},$$

*and*

$$\left\|\hat{\theta}_{2,\tau} - \theta_{\star,2}\right\|_{\mathsf{F}} \leq A\bar{\kappa}\|\theta_{\star,2}\|_2^2\sqrt{\frac{s_2\log(pT)}{T-\tau}},$$

*for all $\tau\in\mathcal{T}$, where $A$ is an absolute constant that can be taken as $A = 16\times 20\times\sqrt{48}$.*

*Proof.* Fix $j\in\{1,2\}$, and $\tau\in\mathcal{T}$. Set $\bar{g}_{j,\tau}(\theta) \stackrel{\text{def}}{=} g_{j,\tau}(\theta) + (1-\alpha)\lambda_{j,\tau}\|\theta\|_{\mathsf{F}}/2$, and recall that $\phi_{j,\tau}(\theta) \stackrel{\text{def}}{=} g_{j,\tau}(\theta) + \lambda_{j,\tau}\wp(\theta)$. Hence $\phi_{j,\tau}(\theta) = \bar{g}_{j,\tau}(\theta) + \alpha\lambda_{j,\tau}\|\theta\|_1$. By a very standard argument that can be found for instance in Negahban et al. (2012), it is known that on the event $\mathcal{E}_n$, and if $\alpha$ satisfies (9) then we have $\hat{\theta}_{j,\tau} - \theta_{\star,j}\in\mathcal{C}_j$, where the cones $\mathcal{C}_j$ are as defined in (16). We write

$$\begin{aligned}\phi_{j,\tau}(\hat{\theta}_{j,\tau}) - \phi_{j,\tau}(\theta_{\star,j}) &= \left\langle\nabla g_{j,\tau}(\theta_{\star,j}) + (1-\alpha)\lambda_{j,\tau}\theta_{\star,j}, \hat{\theta}_{j,\tau} - \theta_{\star,j}\right\rangle\\ &\quad + \bar{g}_{j,\tau}(\hat{\theta}_{j,\tau}) - \bar{g}_{j,\tau}(\theta_{\star,j}) - \left\langle\nabla\bar{g}_{j,\tau}(\theta_{\star,j}), \hat{\theta}_{j,\tau} - \theta_{\star,j}\right\rangle\\ &\quad + \alpha\lambda_{j,\tau}\left(\|\hat{\theta}_{j,\tau}\|_1 - \|\theta_{\star,j}\|_1\right).\end{aligned}$$

On $\mathcal{E}_n$, $\hat{\theta}_{j,\tau} - \theta_{\star,j}\in\mathcal{C}_j$. Therefore

$$\alpha\lambda_{j,\tau}\left|\|\hat{\theta}_{j,\tau}\|_1 - \|\theta_{\star,j}\|_1\right| \leq \alpha\lambda_{j,\tau}\left\|\hat{\theta}_{j,\tau} - \theta_{\star,j}\right\|_1 \leq 8\alpha\lambda_{j,\tau}\sqrt{s_j}\left\|\hat{\theta}_{j,\tau} - \theta_{\star,j}\right\|_{\mathsf{F}},$$

and

$$\begin{aligned}\left|\left\langle \nabla g_{j,\tau}(\theta_{\star,j}) + (1-\alpha)\lambda_{j,\tau}\theta_{\star,j}, \hat{\theta}_{j,\tau} - \theta_{\star,j}\right\rangle\right| \\ \leq \frac{\lambda_{j,\tau}}{2}\left(\alpha + 2(1-\alpha)\|\theta_{\star,j}\|_\infty\right)\left\|\hat{\theta}_{j,\tau} - \theta_{\star,j}\right\|_1 \\ \leq 4\lambda_{j,\tau}\left(\alpha + 2(1-\alpha)\|\theta_{\star,j}\|_\infty\right)\sqrt{s_j}\left\|\hat{\theta}_{j,\tau} - \theta_{\star,j}\right\|_{\mathsf{F}}.\end{aligned}$$

Suppose $j = 1$. The case $j = 2$ is similar. We then set $\Delta_{1,\tau} \stackrel{\text{def}}{=} \hat{\theta}_{1,\tau} - \theta_{\star,1}$, and use the second part of Lemma 12 (1) to deduce that

$$\begin{aligned}\bar{g}_{1,\tau}(\hat{\theta}_{1,\tau}) - \bar{g}_{1,\tau}(\theta_{\star,1}) - \left\langle \nabla \bar{g}_{1,\tau}(\theta_{\star,1}), \hat{\theta}_{1,\tau} - \theta_{\star,1}\right\rangle \\ \geq g_{1,\tau}(\hat{\theta}_{1,\tau}) - g_{1,\tau}(\theta_{\star,1}) - \left\langle \nabla g_{1,\tau}(\theta_{\star,1}), \hat{\theta}_{1,\tau} - \theta_{\star,1}\right\rangle \\ \geq \frac{\tau}{2T}\frac{\|\Delta_{1,\tau}\|_{\mathsf{F}}^2}{2\|\theta_{\star,1}\|_2\left(2\|\theta_{\star,1}\|_2 + \|\Delta_{1,\tau}\|_{\mathsf{F}}\right)}.\end{aligned}$$

Set $c_1 = \frac{\tau}{4T\|\theta_{\star,1}\|_2^2}$, $c_2 = 4\lambda_{1,\tau}\sqrt{s_1}\left(3\alpha + 2(1-\alpha)\|\theta_{\star,1}\|_\infty\right)$. Since $\phi_{1,\tau}(\hat{\theta}_{1,\tau}) - \phi_{1,\tau}(\theta_{\star,1}) \leq 0$, the above derivation shows that on the event $\mathcal{E}_n$,

$$\frac{c_1\|\Delta_{1,\tau}\|_{\mathsf{F}}^2}{2 + \frac{1}{\|\theta_{\star,1}\|_2}\|\Delta_{1,\tau}\|_{\mathsf{F}}} - c_2\|\Delta_{1,\tau}\|_{\mathsf{F}} \leq 0,$$

Under the assumption that $c_1 \geq 2c_2/\|\theta_{\star,1}\|_2$ (which we impose in (10)), this implies that

$$\|\Delta_{1,\tau}\|_{\mathsf{F}} \leq \frac{4c_2}{c_1} \leq A\bar{\kappa}\|\theta_{\star,1}\|_2^2\sqrt{\frac{s_1\log(pT)}{\tau}},$$

where $A = 16 \times 20 \times \sqrt{48}$, as claimed. □

*Proof of Theorem 9.* For $\tau \in \mathcal{T}$, let

$$r_{1,\tau} \stackrel{\text{def}}{=} A\bar{\kappa}\|\theta_{\star,1}\|_2^2\sqrt{\frac{s_1\log(pT)}{\tau}}, \quad r_{2,\tau} \stackrel{\text{def}}{=} A\bar{\kappa}\|\theta_{\star,2}\|_2^2\sqrt{\frac{s_2\log(pT)}{T-\tau}},$$

be the convergence rates obtained in Lemma 16. Let $\epsilon > 0$ be given by

$$\epsilon \stackrel{\text{def}}{=} \min_{\tau\in\mathcal{T}}(r_{1,\tau} \wedge r_{1,\tau}).$$

For $j = 1, 2$, let $\theta_j \in \mathcal{M}_p^+$ be such that $\|\theta_j - \hat{\theta}_{\tau,j}\|_1 \leq \epsilon$. Set $\check{\tau} = \mathsf{Argmin}_{t\in\mathcal{T}}\,\mathcal{H}(t|\theta_1,\theta_2)$, where $\mathcal{H}$ is as defined in (4). Set

$$C_0 = \min\left[\frac{\|\theta_{\star,2} - \theta_{\star,1}\|_{\mathsf{F}}^4}{128B^4\|\theta_{\star,2} - \theta_{\star,1}\|_1^2}, \left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^4\right].$$

We will show below that

$$\mathbb{P}\left(|\check{\tau} - \tau_\star| > \frac{4\log(p)}{C_0}\right) \leq \frac{8}{pT} + \frac{4}{p^2\left(1 - e^{-C_0}\right)}. \tag{18}$$

This implies that with probability at least $1 - \frac{8}{pT} - \frac{4}{p^2\left(1-e^{-C_0}\right)}$, Assumption H1 holds (with $\epsilon \leftarrow \epsilon/\sqrt{p}$, $\kappa = 0$, and $c = (4/C_0)\log(p)$). The theorem then follows by applying Theorem 5.

Given $\theta_j \in \mathcal{M}_p^+$ be such that $\|\theta_j - \hat{\theta}_{\tau,j}\|_1 \leq \epsilon$, we will now show that (18) holds. We shall bound $\mathbb{P}(\check{\tau} > \tau_\star + \delta)$, $\delta = (4/C_0)\log(p)$. The bound on $\mathbb{P}(\check{\tau} < \tau_\star - \delta)$ follows similarly by working with the reversed sequence $X^{(T)}, \ldots, X^{(1)}$.

Note that $\theta_j$ can be written as

$$\theta_j = (\theta_j - \hat{\theta}_{\tau,j}) + (\hat{\theta}_{\tau,j} - \theta_{\star,j}) + \theta_{\star,j}. \tag{19}$$

This implies that on $\mathcal{E}_n$, for $\epsilon \leq r_{j,\tau}$, and $r_{j,\tau} \leq \min\left(\frac{\lambda_{\mathsf{min}}(\theta_{\star,j})}{4}, \frac{\|\theta_{\star,j}\|_\infty}{2}, \frac{\|\theta_{\star,j}\|_1}{1+8s_j^{1/2}}\right)$, we have

$$\lambda_{\mathsf{min}}(\theta_j) \geq \frac{1}{2}\lambda_{\mathsf{min}}(\theta_{\star,j}), \quad \lambda_{\mathsf{max}}(\theta_j) \leq 2\lambda_{\mathsf{max}}(\theta_{\star,j}),$$
$$\|\theta_j\|_\infty \leq 2\|\theta_{\star,j}\|_\infty, \quad \text{and} \quad \|\theta_j\|_1 \leq 2\|\theta_{\star,j}\|_1. \tag{20}$$

Using the event $\mathcal{E}_n$ introduced in (17), we have

$$\begin{aligned}\mathbb{P}\left(\check{\tau} > \tau_\star + \delta\right) &\leq \mathbb{P}(\mathcal{E}_n^c) + \sum_{j\geq 0:\ \tau_\star+\delta+j\in\mathcal{T}} \mathbb{P}\left(\mathcal{E}_n,\ \check{\tau} = \tau_\star + \delta + j\right) \\ &\leq \mathbb{P}(\mathcal{E}_n^c) + \sum_{j\geq 0:\ \tau_\star+\delta+j\in\mathcal{T}} \mathbb{P}\left(\mathcal{E}_n,\ \phi_{1,\tau_\star+\delta+j}(\theta_1) + \phi_{2,\tau_\star+\delta+j}(\theta_2) \leq \phi_{1,\tau_\star}(\theta_1) + \phi_{2,\tau_\star}(\theta_2)\right),\end{aligned} \tag{21}$$

where $\phi_{j,\tau}(\theta) \stackrel{\text{def}}{=} g_{j,\tau}(\theta) + \lambda_{j,\tau}\wp(\theta)$. First we are going to bound the probability

$$\mathbb{P}\left(\mathcal{E}_n,\ \phi_{1,\tau}(\theta_1) + \phi_{2,\tau}(\theta_2) \leq \phi_{1,\tau_\star}(\theta_1) + \phi_{2,\tau_\star}(\theta_2)\right),$$

for some arbitrary $\tau \in \mathcal{T}$, $\tau > \tau_\star$. A simple calculation shows that

$$\begin{aligned}\frac{2T}{\tau - \tau_\star}\left[\phi_{1,\tau}(\theta_1) + \phi_{2,,\tau}(\theta_2) - \phi_{1,\tau_\star}(\theta_1) - \phi_{2,\tau_\star}(\theta_2)\right] &= -\log\det(\theta_1) + \log\det(\theta_2) \\ &+ \left\langle \theta_1 - \theta_2, \theta_{\star,2}^{-1}\right\rangle + \left\langle \theta_1 - \theta_2, \frac{1}{\tau - \tau_\star}\sum_{t=\tau_\star+1}^{\tau}\left(X^{(t)}X^{(t)\prime} - \theta_{\star,2}^{-1}\right)\right\rangle \\ &+ 2T\left(\frac{\lambda_{1,\tau} - \lambda_{1,\tau_\star}}{\tau - \tau_\star}\right)\left(\frac{1-\alpha}{2}\|\theta_1\|_{\mathsf{F}}^2 + \alpha\|\theta_1\|_1\right) \\ &+ 2T\left(\frac{\lambda_{2,\tau} - \lambda_{2,\tau_\star}}{\tau - \tau_\star}\right)\left(\frac{1-\alpha}{2}\|\theta_2\|_{\mathsf{F}}^2 + \alpha\|\theta_2\|_1\right).\end{aligned}$$

We have $2T\left(\frac{\lambda_{1,\tau}-\lambda_{1,\tau_\star}}{\tau-\tau_\star}\right)\left(\frac{1-\alpha}{2}\|\theta_1\|_{\mathsf{F}}^2+\alpha\|\theta_1\|_1\right)\geq 0$, and

$$2T\left|\frac{\lambda_{2,\tau}-\lambda_{2,\tau_\star}}{\tau-\tau_\star}\right|\leq\frac{\bar{\kappa}}{\alpha}\sqrt{\frac{48\log(pT)}{T-\tau}}=\frac{c_0 r_{2,\tau}}{\alpha s_2^{1/2}\|\theta_{\star,2}\|_2^2},$$

for some absolute constant $c_0$. Using the infinity-norm and 1-norm bounds in (20) together with (9), we have

$$\frac{1-\alpha}{2}\|\theta_2\|_{\mathsf{F}}^2+\alpha\|\theta_2\|_1=\alpha\left[\frac{1-\alpha}{2\alpha}\|\theta_2\|_\infty+1\right]\|\theta_2\|_1\leq 4\alpha\|\theta_{\star,2}\|_1,$$

and it follows that

$$2T\left|\frac{\lambda_{2,\tau}-\lambda_{2,\tau_\star}}{\tau-\tau_\star}\right|\left(\frac{1-\alpha}{2}\|\theta_2\|_{\mathsf{F}}^2+\alpha\|\theta_2\|_1\right)\leq C_\tau\stackrel{\text{def}}{=}\left(\frac{4c_0\|\theta_{\star,2}\|_1}{s_2^{1/2}\|\theta_{\star,2}\|_2^2}\right)r_{2,\tau}.$$

Set

$$b\stackrel{\text{def}}{=}\min\left(\lambda_{\mathsf{min}}(\theta_{\star,1}),\lambda_{\mathsf{min}}(\theta_{\star,2})\right),\quad B\stackrel{\text{def}}{=}\max\left(\|\theta_{\star,1}\|_2,\|\theta_{\star,2}\|_2\right).$$

By the strong convexity of $\log\det$ (Lemma 12 Part(1)) we have:

$$\begin{aligned}-\log\det(\theta_1)+\log\det(\theta_2)+\left\langle\theta_1-\theta_2,\theta_{\star,2}^{-1}\right\rangle&\\ \geq\left\langle\theta_{\star,2}^{-1}-\theta_2^{-1},\theta_1-\theta_2\right\rangle+\frac{1}{2B^2}\|\theta_1-\theta_2\|_{\mathsf{F}}^2.&\end{aligned}$$

Since $\theta_{\star,2}^{-1}-\theta_2^{-1}=\theta_{\star,2}^{-1}(\theta_2-\theta_{\star,2})\theta_2^{-1}$, and using the fact that $\|AB\|_{\mathsf{F}}\leq\|A\|_2\|B\|_{\mathsf{F}}$, we have that on $\mathcal{E}_n$,

$$\left|\left\langle\theta_{\star,2}^{-1}-\theta_2^{-1},\theta_1-\theta_2\right\rangle\right|\leq 2r_{2,\tau}\|\theta_{\star,2}^{-1}\|_2\|\theta_2^{-1}\|_2\|\theta_2-\theta_1\|_{\mathsf{F}}\leq 4r_{2,\tau}\|\theta_{\star,2}^{-1}\|_2^2\|\theta_2-\theta_1\|_{\mathsf{F}}.$$

We conclude that on $\mathcal{E}_n$,

$$\begin{aligned}\frac{2T}{\tau-\tau_\star}\left[\phi_{1,\tau}(\theta_1)+\phi_{2,\tau}(\theta_2)-\phi_{1,\tau_\star}(\theta_1)-\phi_{2,\tau_\star}(\theta_2)\right]\geq&\\ \left\langle\theta_1-\theta_2,\frac{1}{\tau-\tau_\star}\sum_{t=\tau_\star+1}^{\tau}\left(X^{(t)}X^{(t)\prime}-\theta_{\star,2}^{-1}\right)\right\rangle&\\ -C_\tau-4r_{2,\tau}\|\theta_{\star,2}^{-1}\|_2^2\|\theta_2-\theta_1\|_{\mathsf{F}}+\frac{1}{2B^2}\|\theta_1-\theta_2\|_{\mathsf{F}}^2.&\end{aligned}$$

Under the assumption (12) imposed on $r_{j,\tau}$ and for $\epsilon\leq r_{1,\tau}\wedge r_{2,\tau}$, it can be shown that on $\mathcal{E}_n$, and for $\|\theta_{\star,2}-\theta_{\star,1}\|_{\mathsf{F}}\geq\frac{8c_0\|\theta_{\star,2}\|_1}{s_2^{1/2}\|\theta_{\star,2}\|_2^2\|\theta_{\star,2}^{-1}\|_2^2}$, we have

$$-C_\tau-2\left(\epsilon+r_{2,\tau}\right)\|\theta_{\star,2}^{-1}\|_2^2\|\theta_2-\theta_1\|_{\mathsf{F}}+\frac{1}{4B^2}\|\theta_1-\theta_2\|_{\mathsf{F}}^2\geq 0.\tag{22}$$

To see this, note that (22) holds if $\|\theta_2-\theta_1\|_{\mathsf{F}} \geq 8B^2 r_{2,\tau}\|\theta_{\star,2}^{-1}\|_2^2 + 2B\sqrt{C_\tau + 16B^2\|\theta_{\star,2}^{-1}\|_2^4 r_{2,\tau}^2}$. Then it can be checked that if $r_{2,\tau} \leq \frac{c_0\|\theta_{\star,2}\|_1}{16B^2 s_2^{1/2}\|\theta_{\star,2}\|_2^2\|\theta_{\star,2}^{-1}\|_2^2}$, then

$$8B^2\|\theta_{\star,2}^{-1}\|_2^2 r_{2,\tau} \leq \frac{C_\tau}{2\|\theta_{\star,2}^{-1}\|_2^2 r_{2,\tau}}, \quad \text{and} \quad 4B\sqrt{C_\tau} \leq \frac{C_\tau}{2\|\theta_{\star,2}^{-1}\|_2^2 r_{2,\tau}}.$$

Therefore, (22) holds if

$$\|\theta_2-\theta_1\|_{\mathsf{F}} \geq \frac{C_\tau}{\|\theta_{\star,2}^{-1}\|_2^2 r_{2,\tau}} = \frac{4c_0\|\theta_{\star,2}\|_1}{s_2^{1/2}\|\theta_{\star,2}\|_2^2\|\theta_{\star,2}^{-1}\|_2^2}.$$

Now we write

$$\theta_2-\theta_1 = (\theta_2-\hat{\theta}_{\tau,2}) + (\hat{\theta}_{\tau,2}-\theta_{\star,2}) + (\theta_{\star,2}-\theta_{\star,1}) + (\theta_{\star,1}-\hat{\theta}_{\tau,1}) + (\hat{\theta}_{\tau,1}-\theta_1),$$

and use the fact that $\epsilon \leq r_{1,\tau}\wedge r_{2,\tau}$, and $r_{j,\tau} \leq \|\theta_{\star,2}-\theta_{\star,1}\|_{\mathsf{F}}/8$ to deduce that on $\mathcal{E}_n$, $\|\theta_2-\theta_1\|_{\mathsf{F}} \geq \|\theta_{\star,2}-\theta_{\star,1}\|_{\mathsf{F}}/2$, and this completes the proof of the claim.

It follows from the above that

$$\mathbb{P}\left(\mathcal{E}_n; \phi_{1,\tau}(\theta_1)+\phi_{2,\tau}(\theta_2)-\phi_{1,\tau_\star}(\theta_1)-\phi_{2,\tau_\star}(\theta_2) \leq 0\right)$$
$$\leq \mathbb{P}\left(\left\|\frac{1}{\tau-\tau_\star}\sum_{t=\tau_\star+1}^{\tau}\left(X^{(t)}X^{(t)\prime}-\theta_{\star,2}^{-1}\right)\right\|_\infty > \frac{\|\theta_2-\theta_1\|_{\mathsf{F}}^2}{4B^2\|\theta_2-\theta_1\|_1}\right). \quad (23)$$

Proceeding as above, it is easy to see that if $\epsilon \leq r_{1,\tau}\wedge r_{2,\tau}$, and $r_{j,\tau} \leq \frac{\|\theta_{\star,2}-\theta_{\star,1}\|_{\mathsf{F}}}{2(1+8s^{1/2})}$, then

$$\frac{\|\theta_2-\theta_1\|_{\mathsf{F}}^2}{4B^2\|\theta_2-\theta_1\|_1} \geq \frac{\|\theta_{\star,2}-\theta_{\star,1}\|_{\mathsf{F}}^2}{32B^2\|\theta_{\star,2}-\theta_{\star,1}\|_1}.$$

Using this, and by Lemma 15, it follows that the probability on the right-hand side of (23) is upper-bounded by

$$4p^2\exp\left(-(\tau-\tau_\star)\min\left[\frac{\|\theta_{\star,2}-\theta_{\star,1}\|_{\mathsf{F}}^4}{128B^4\|\theta_{\star,2}-\theta_{\star,1}\|_1^2}, \left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^4\right]\right).$$

We apply this to (21) to get:

$$\mathbb{P}(\check{\tau} > \tau_\star+\delta) \leq \mathbb{P}(\mathcal{E}_n^c) + \sum_{j\geq 0} 4p^2 e^{-C_0(\delta+j)} \leq \frac{8}{pT} + \frac{4}{p^2(1-e^{-C_0})},$$

where $C_0 = \min\left[\frac{\|\theta_{\star,2}-\theta_{\star,1}\|_{\mathsf{F}}^4}{128B^4\|\theta_{\star,2}-\theta_{\star,1}\|_1^2}, \left(\frac{\underline{\kappa}}{\bar{\kappa}}\right)^4\right]$, and by taking $\delta = 4\log(p)/C_0$. This completes the proof.

□

## References


Atchade, Y. F., Fort, G. and Moulines, E. (2017). On stochastic proximal gradient algorithms. *Journal of Machine Learning Research* **18** 1–33.

Atchadé, Y. F., Mazumder, R. and Chen, J. (2015). Scalable Computation of Regularized Precision Matrices via Stochastic Optimization. *ArXiv e-prints* .

Aue, A., Hormann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *Ann. Statist.* **37** 4046–4087.

Bai, J. (1997). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics* **79** 551–563.

Banerjee, A. and Urga, G. (2005). Modelling structural breaks, long memory and stock market volatility: an overview. *Journal of Econometrics* **129** 1–34.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516.

Beltratti, A. and Morana, C. (2006). Breaks and persistency: macroeconomic causes of stock market volatility. *Journal of econometrics* **131** 151–177.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604.

Bybee, L. (2017). *changepointsHD: Change-Point Estimation for Expensive and High-Dimensional Models*. R package version 0.3.0.

Chen, H. and Zhang, N. (2015). Graph-based change-point detection. *Ann. Statist.* **43** 139–176.

Cho, H. and Fryzlewicz, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 475–507.

Choi, K., Yu, W.-C. and Zivot, E. (2010). Long memory versus structural breaks in modeling and forecasting realized volatility. *Journal of International Money and Finance* **29** 857–875.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281.

Günay, S. (2014). Long memory property and structural breaks in volatility: Evidence from turkey and brazil. *International Journal of Economics and Finance* **6** 119.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906.

Kolar, M., Song, L., Ahmed, A. and Xing, E. (2010). Estimating time-varying networks. *Ann. Appl. Statist.* **4** 94–123.

Kolar, M. and Xing, E. P. (2012). Estimating networks with jumps. *Electron. J. Statist.* **6** 2069–2106.

Lafferty, J., Liu, H., Wasserman, L. et al. (2012). Sparse nonparametric graphical models. *Statistical Science* **27** 519–537.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338.

Leonardi, F. and Bühlmann, P. (2016). Computationally efficient change point detection for high-dimensional regression. *ArXiv e-prints* .

Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data. *Ann. Appl. Stat.* **3** 637–662.

Liu, S., Quinn, J. A., Gutmann, M. U. and Sugiyama, M. (2013). Direct learning of sparse changes in markov networks by density ratio estimation. In *Machine Learning and Knowledge Discovery in Databases* (H. Blockeel, K. Kersting, S. Nijssen and F. Železný, eds.). Springer Berlin Heidelberg, Berlin, Heidelberg.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs with the lasso. *Annals of Stat.* **34** 1436–1462.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of $m$-estimators with decomposable regularizers. *Statistical Science* **27** 538–557.

Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* **1** 123–231.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. *Ann. Statist.* **38** 1287–1319.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing $\ell_1$-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980.

Rolfs, B., Rajaratnam, B., Guillot, D., Wong, I. and Maleki, A. (2012). Iterative thresholding algorithm for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.). Curran Associates, Inc., 1574–1582.

Roy, S., Atchadé, Y. and Michailidis, G. (2017). Change point estimation in high dimensional markov random-field models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1187–1206.

Wu, T. T. and Lange, K. (2010). The mm alternative to em. *Statist. Sci.* **25** 492–505.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35.

Zhou, S., Lafferty, J. and Wasserman, L. (2009). Time-varying undirected graphs. *In Rocco A. Dervedio and Tong Zhang, editors, Conference on learning Theory* 455–466.